STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

ZW 1949-013

S.O. 1

Over de bepaling van betrouwbaarheidsintervallen en schattingen van de coëfficiënten van een rechte lijn uit een aantal onnnauwkeurig waargenomen punten

J. Hemelrijk

1949

Errata bij Rapport S.O. 1.

bladzijde en regel:			
4	15 v.b.	door	door voor
7	15 v.o.	achter "indien" in te voegen: voor iedere i.	
10	17 v.b.	aan het eind van voorwaarde 4) toevoegen: $= 0$.	
	4 v.o.	$S_{\xi\eta}$	$S_{\xi\eta}$
16	11 v.o.	4)	3)
35	13 v.b.	$\frac{1}{\alpha}$	ξ_0
38	1 v.b.	nu	men
	19 v.o.	methodes	methods
	9 v.o.	se	use

Rapport S.9.'f(1949)
tevens ZW-(1949)-013.

Over de bepaling van betrouwbaarheidsintervallen en schattingen van de coëfficienten van een rechte lijn uit een aantal onnauwkeurig waargenomen punten.

Overzichtsrapport no. 1.

Door J. Hemelrijk.

Inhoud.

I. Inleiding. pag.

1. Het probleem 2
2. Waarschijnlijkheidstheoretische interpretatie; 2 gevallen 3
3. Schattingen en betrouwbaarheidsintervallen (definities) 4

II. Verschillende oplossingen.

4. Methode der kleinste quadraten en maximum-likelihoodmethode 6
5. Methode van A. Wald 12
6. Methode van G.W. Housner en J.F. Brennan . . . 15
7. Onzuiverheid van deze methoden 17
8. Betrouwbaarheidsgrenzen voor de coëfficienten zonder onderstelling van normaliteit van de fouten 19
9. Methoden van H. Theil 24
10. Het geval, dat ξ een waarschijnlijkheidsverdeling bezit 27

III. Schatting van één coördinaat als de andere gegeven is.

11. Het probleem 32
12. Schatting van de abscis bij gegeven ordinaat, als de abscis foutloos is 33
13. Bruikbaarheid van schattingen van de ene coördinaat uit de andere 35
14. Keuze uit de verschillende methoden 36
15. Generalisatie van de theorie 38
16. Lijst van geciteerde litteratuur . . . . . 38.

I. Inleiding.

1. Het probleem.

1.1. Wij onderstellen de volgende situatie:

1) ξ en η zijn twee grootheden, die naar men weet (of aanneemt) door een lineaire betrekking L met onbekende coëfficiënten met elkaar verbonden zijn:

(1) $L: \eta = \alpha\xi + \beta \qquad -\infty \leq \alpha \leq +\infty \qquad -\infty \leq \beta \leq +\infty$

(Voorbeelden: ξ en η zijn dezelfde grootheid, gemeten met verschillende schalen, zodat L een ijklijn is; ξ is een stroomsterkte, η een spanning, α een onbekende weerstand, $\beta = 0$; bij een chemische reactie: ξ is de concentratie van één der reagentia, η de reactiesnelheid, terwijl α door de overige, onbekende, concentraties wordt bepaald, $\beta = 0$). De lijn $\eta = \alpha\xi + \beta$ geven wij met L aan.

2) Waarneming van punten van L gaat gepaard met waarnemingsfouten in één van beide of beide grootheden. Als waarnemingsresultaat vindt men dus in plaats van een punt Q van L met coördinaten (ξ, η) een punt P met coördinaten (x,y), waarbij in het algemeen $x \neq \xi$ en/of $y \neq \eta$ is.

3) Gegeven zijn n dergelijke waargenomen punten $P_i \equiv (x_i, y_i)$ $(i = 1,\ldots,n)$, behorend bij onbekende punten Q_i van L, waarvan er minstens twee verschillend zijn.

Gevraagd wordt, wat, op grond van waarschijnlijkheidtheoretische overwegingen en met behulp van later te preciseren onderstellingen omtrent de waarschijnlijkheidsverdelingen van de meetfouten, van de α en β gezegd kan worden.

1.2. In het bijzonder zullen wij verschillende schattingsmethoden voor α en β beschouwen en enkele methoden, die het mogelijk maken, voor α en β betrouwbaarheidsintervallen te bepalen. Tenslotte beschouwen wij nog het probleem van de schatting van één van de twee coördinaten van een punt Q (resp. P) als de andere coördinaat gegeven is (met of zonder meetfout).

Het overzicht, dat in deze syllabus gegeven wordt, is niet volledig; slechts de nieuwere en de meest bruikbare van de oudere methoden worden besproken. De bewijzen zijn in het algemeen slechts schematisch aangegeven, met vermelding van de plaats, waar zij volledig te vinden zijn.

Een historisch overzicht over de ontwikkeling van dit probleem vindt men bij A. WALD [17] en D.V. LINDLEY [12].

2. Waarschijnlijkheidstheoretische interpretatie; 2 gevallen.

2.1. Notaties. De elementaire begrippen van de mathematische statistiek worden bekend ondersteld.

De notatie van Prof. Dr D. van DANTZIG volgende, geven wij het stochastisch karakter van een stochastische variabele (algemeen van een stochastisch element van een verzameling) aan door het bijbehorende symbool te onderstrepen. [1]) Parameters (ook onbekende), die een waarschijnlijkheidsverdeling bepalen, worden met Griekse letters aangegeven en bepalende parameters genoemd, stochastische variabelen worden met Latijnse letters aangegeven.

De kans op een gebeurtenis A, onder de voorwaarde B, wordt aangegeven door $P_B[A]$ of $P[A|B]$; dezelfde notatie wordt gebruikt als B een hypothese voorstelt; de onvoorwaardelijke waarschijnlijkheid van A door $P[A]$. de mathematische verwachting van een stochastische variabele wordt aangegeven door het operatiesymbool $\mathcal{E}$; voorwaarden worden op dezelfde wijze aangeduid als bij P.

De waargenomen punten $P_1,\ldots,P_n$ zullen wij tezamen één waarnemingsresultaat noemen, behorende bij de met $P_1,\ldots,P_n$ corresponderende punten $Q_1,\ldots,Q_n$ van L. Q_i zal soms het bij P_i behorende "ware" punt worden genoemd. Noemen wij de meetfouten in ξ - en η -richting u_i resp. v_i, dan hebben wij:

$$(2)\qquad \begin{cases} x_i = \xi_i + u_i \qquad y_i = \eta_i + v_i \qquad i = 1,\ldots,n \\ \eta_i = \alpha\xi_i + \beta \end{cases}$$

2.2. Voor de waarschijnlijkheidstheoretische interpretatie onderscheiden wij twee gevallen:

Geval I. De grootheid ξ bezit geen waarschijnlijkheidsverdeling (of deze bestaat wel, maar wordt buiten beschouwing gelaten). Wegens (1) is ditzelfde dan voor η het geval. De waarschijnlijkheidstheoretische overwegingen berusten in dit geval op de overgang van het gevonden waarnemingsresultaat naar de collectie ⌜ van alle mogelijke bij dezelfde punten $Q_1,\ldots,Q_n$ behorende waarnemingsresultaten. De onderstellingen, die omtrent de meetfouten gemaakt zullen worden maken ⌜ tot een waarschijnlijkheidsveld.

---

[1]) Een stochastische variabele is een variabele, die een waarschijnlijkheidsverdeling bezit.

Indien de onderstelling gemaakt wordt, dat de paren meetfouten $(\underline{u}_i, \underline{v}_i)$ onderling onafhankelijk verdeeld zijn, kan men Γ opgebouwd denken als productveld van n twee-dimensionale waarschijnlijkheidsvelden Γ_i $(i = 1,\ldots,n)$, waarbij Γ_i het waarschijnlijkheidsveld is behorende bij het met Q_i corresponderend stochastische punt $\underline{P}_i$.

Deze waarschijnlijkheidsinterpretatie wordt aangegeven door de vergelijkingen (2) te schrijven in de vorm:

$$(2') \qquad \underline{x}_i = \xi_i + \underline{u}_i \qquad \underline{y}_i = \eta_i + \underline{v}_i \qquad i = 1,\ldots,n$$

verbonden door de lineaire betrekking:

$$(1') \qquad \eta_i = \alpha \xi_i + \beta \qquad i = 1,\ldots,n$$

met ξ_i, α en β als onbekende parameters, terwijl ook de verdelingen van $\underline{u}_i$ en $\underline{v}_i$ nog onbekende parameters kunnen bevatten.

Geval II. De grootheid ξ bezit wèl een waarschijnlijkheidsverdeling (dus η ook); wij zullen dit aangeven door ξ en η te schrijven $\underline{X}$ en $\underline{Y}$ (zodat $\underline{Y} = \alpha \underline{X} + \beta$ geldt). Het punt $\underline{Q} = (\underline{X}, \underline{Y})$ is dan stochastisch verdeeld over L en het probleem gaat over in een speciaal geval van de tweedimensionale regressieanalyse; (2) gaat over in:

$$(2'') \qquad \underline{x}_i = \underline{X}_i + \underline{u}_i \qquad \underline{y}_i = \underline{Y}_i + \underline{v}_i \qquad i = 1,\ldots,n$$

waarin $\underline{X}_i$ en $\underline{Y}_i$ verbonden zijn door de lineaire betrekking:

$$(1'') \qquad \underline{Y}_i = \alpha \underline{X}_i + \beta \qquad i = 1,\ldots,n$$

met α en β als onbekende parameters, terwijl ook de verdelingen van $\underline{u}_i$, $\underline{v}_i$ en $\underline{X}_i$ nog onbekende parameters kunnen bevatten.

Het met deze situatie corresponderence wh-veld Γ^* is de collectie van alle mogelijke waarnemingsresultaten van de uitgebreidheid n, waarbij de punten $\underline{Q}_i$ nu echter onafhankelijk van elkaar verdeeld zijn met een, van i onafhankelijke waarschijnlijkheidsverdeling, die uit de boven onderstelde waarschijnlijkheidsverdeling van $\underline{X}$ voortvloeit.

Wij zullen ons in hoofdzaak bezig houden met het eerste geval; aan het slot echter komen wij op geval II terug.

3. Schattingen en betrouwbaarheidsintervallen (definities).

3.1. Iedere meetbare functie $\underline{t} = t(\underline{x}_1,\ldots,\underline{x}_n; \underline{y}_1,\ldots,\underline{y}_n)$ van de coördiraten van de stochastische punten $\underline{P}_i$ bezit een waarschijnlijkheidsverdeling, die van de onbekende parameters afhankelijk is. Een dergelijke functie wordt, indien zij zelf

niet van onbekende paramters afhangt, een statistische variabele genoemd.

Een statistische variabele wordt een schatting van een onbekende parameter θ genoemd, indien haar verdeling aan bepaalde eisen voldoet. Als minimum-eis stelt men gewoonlijk, dat $\underline{t}$ de eerste der volgende eigenschappen moet bezitten (de vertaling der Engelse termen is ontleend aan D. van DANTZIG [3] p. 202 (Whr 291)):

a) $\underline{t}$ heet een bruikbare (Engels: consistent) schatting van θ, als

$$(3) \quad \lim_{n\to\infty} P[|\underline{t}_n - \theta| < \varepsilon \mid \theta] = 1 \qquad \text{voor iedere } \varepsilon > 0 .$$

In woorden: indien $\underline{t}_n$ de schatting $\underline{t}$ voorstelt bij uitgebreidheid n van de waarnemingsreeks (of in het algemeen: steekproef) en θ de ware waarde van de bepalende parameter is, nadert de kans, dat $\underline{t}_n$ minder dan ε van θ verschilt, voor iedere constante $\varepsilon > 0$, tot 1 als n naar ∞ gaat.

Aequivalent met (3) is

Bij iedere $\varepsilon > 0$ en $\delta > 0$ is er een natuurlijk getal $N(\varepsilon, \delta)$, dat van ε en δ afhangt met

$$(4) \quad P[|\underline{t}_n - \theta| < \varepsilon] > 1 - \delta \qquad \text{voor iedere } n > N(\varepsilon, \delta) .$$

Men zegt dan, dat $\underline{t}_n$ stochastisch naar θ convergeert.

b) $\underline{t}$ heet een zuivere (Engels: unbiased) schatting van θ als voor iedere uitgebreidheid n geldt:

$$(5) \quad \mathcal{E}_\theta \underline{t}_n = \theta$$

d.w.z. de verwachting van $\underline{t}_n$, is gelijk aan de ware parameterwaarde is, (i.c. θ).

c) $\underline{t}$ heet de doeltreffendste (Engels: most efficient) schatting van θ, als $\underline{t}$ een zuivere schatting is, die bovendien van alle zuivere schattingen de kleinste spreiding bezit. D.w.z. als voor iedere andere zuivere schatting $\underline{t}^*_n$ van θ geldt:

$$(6) \quad \frac{\sigma^2_{\underline{t}_n}}{\sigma^2_{\underline{t}^*_n}} = \frac{\mathcal{E}_\theta\{\underline{t}_n - \mathcal{E}_\theta \underline{t}_n\}^2}{\mathcal{E}_\theta\{\underline{t}^*_n - \mathcal{E}_\theta \underline{t}^*_n\}^2} \geq 1$$

voor iedere n.

Zoals reeds opgemerkt stelt men gewoonlijk aan een schatting de eis van bruikbaarheid; verder is het duidelijk, dat zuiverheid een zeer wenselijke eigenschap is. Bij de keuze tussen mogelijke schattingen laat men zich naast beoordeling van hun

nut naar aanleiding van bovengenoemde eigenschappen, uiteraard tevens leiden door de bewerkelijkheid van hun berekening. Dit laatste echter is uit zuiver wiskundig oogpunt weinig interessant en zal hier buiten beschouwing gelaten worden.

Voor een stelsel statistische variabelen $\underline{t}_1, \ldots, \underline{t}_m$, die een simultane waarschijnlijkheidsverdeling bezitten en beschouwd worden als schattingen van een stelsel bepalende parameters $\theta_1, \ldots, \theta_m$, kan men op analoge wijze bovengenoemde eigenschappen definiëren. Vgl. b.v. D.v.DANTZIG [3] p. 221 en 222 (Whr 310 en 311).

3.2. Twee statistische grootheden $\underline{t}_1$ en $\underline{t}_2$, die een simultane waarschijnlijkheidsverdeling bezitten waarvoor geldt: $\underline{t}_1 < \underline{t}_2$ voor ieder element λ van het waarschijnlijkheidsveld Γ, worden <u>betrouwbaarheidsgrenzen</u> voor een bepalende parameter θ genoemd (of ook: het interval met $\underline{t}_1$ en $\underline{t}_2$ als eindpunten wordt een <u>betrouwbaarheidsinterval</u> voor θ genoemd) indien geldt:

$$(7) \qquad P[\underline{t}_1 < \theta < \underline{t}_2 \mid \theta] = 1 - p \qquad ^{1})$$

waarin p een bekende constante ($0 \leqq p \leqq 1$) is, die de <u>onbetrouwbaarheidsdrempel</u> (Engels: confidence level) van het betrouwbaarheidsinterval heet (of ook: $1-p$ is de <u>betrouwbaarheidsdrempel</u>; Engels: confidence coefficient).

Analoog voor meer dan één parameter; een van de stochastische coördinaten $\underline{x}_1, \ldots, \underline{x}_n$; $\underline{y}_1, \ldots, \underline{y}_n$ afhankelijk gebied $\underline{G}$ in de $(\theta_1, \ldots, \theta_m)$-ruimte is een <u>betrouwbaarheidsgebied</u> voor $(\theta_1, \ldots, \theta_m)$ met betrouwbaarheidsdrempel $1-p$, als voldaan is aan

$$(8) \qquad P[(\theta_1, \ldots, \theta_m) \in \underline{G} \mid \theta_1, \ldots, \theta_m] = 1 - p \quad (resp \geqq 1-p)$$

II. <u>Verschillende oplossingen.</u>

4. <u>Methoden der kleinste quadraten en maximum likelihood methode.</u>

4.1. <u>Beschrijving van de methode der kleinste quadraten.</u>

De oudste methode, waarmee het hier behandelde probleem is aangepakt, is de methode der kleinste quadraten. Een

---

[1]) Indien het niet gelukt een $\underline{t}_1$ en $\underline{t}_2$ te vinden, waarvoor (7) met het gelijkheidsteken geldt, moet men genoegen nemen met een $\geqq$-teken in de plaats daarvan. Bij discrete verdelingen is dit, indien men p van tevoren kiest, veelal noodzakelijk.

nistorische beschouwing daaromtrent vindt men bij D.V. LINDLEY [12] p. 241 e.v.

Lindley beschrijft deze methode als volgt:

Onderstellingen:

1) $\underline{u}_i$ (i = 1,...,n) is normaal verdeeld met gemiddelde 0 en spreiding $\delta/\sqrt{g_i}$, waarin δ onbekend is, maar de g_i bekend zijn.
2) $\underline{v}_i$ (i = 1,...,n) is normaal verdeeld met gemiddelde 0 en spreiding $\delta\sqrt{k}/\sqrt{h_i}$, waarin δ onbekend is, maar k en h_i bekend zijn.
3) $\underline{u}_1, \ldots, \underline{u}_n$; $\underline{v}_1, \ldots, \underline{v}_n$ zijn alle onderling onafhankelijk verdeeld.

Onder deze omstandigheden volgt uit (2'):

$$\sigma^2_{\underline{y}_i - \alpha \underline{x}_i - \beta} = \sigma^2_{\underline{y}_i} + \alpha^2 \sigma^2_{\underline{x}_i} = \delta^2 \left(k/h_i + \alpha^2/g_i\right) = \delta^2 \frac{k g_i + \alpha^2 h_i}{g_i h_i}$$

De methode bestaat nu uit het minimaliseren van de vorm:

$$(9) \qquad \sum i \frac{g_i h_i}{k g_i + \alpha^2 h_i} (y_i - \alpha x_i - \beta)^2$$

hetgeen voor $h_i = g_i$ overgaat in:

$$(9') \qquad \frac{1}{k+\alpha^2} \sum g_i (y_i - \alpha x_i - \beta)^2$$

waarin (x_i, y_i) de coördinaten van de gevonden waarnemingsreeks zijn. Indien $g_i = c.h_i$ is met bekende c kan men dit probleem expliciet oplossen (door wijziging van k kan $c = 1$ gemaakt worden). Door de afgeleiden naar α en β gelijk nul te stellen en enige herleiding toe te passen, verkrijgt men dan als kleinste-quadratenschattingen (die we met een * aangeven): ')

$$(10) \qquad a^{*2} s_{xy} + a^* (k s_x^2 - s_y^2) - k s_{xy} = 0$$

$$(11) \qquad b^* = \bar{y} - a^* \bar{x}$$

waarin

$$(12) \quad \begin{cases} \bar{x} = \frac{1}{n} \sum g_i x_i & \bar{y} = \frac{1}{n} \sum g_i y_i \\ s_x^2 = \frac{1}{n} \sum g_i (x_i - \bar{x})^2 & s_y^2 = \frac{1}{n} \sum g_i (y_i - \bar{y})^2 \\ s_{xy} = \frac{1}{n} \sum g_i (x_i - \bar{x})(y_i - \bar{y}) & \end{cases}$$

is.

') Van grootheden, voorgesteld door Griekse letters, worden de schattingen door de overeenkomstige Latijnse letters voorgesteld.

Uit (10) volgen twee waarden voor a^*; de waarde

(13) $\quad a^* = \zeta + \sqrt{\zeta^2 + k} \qquad met \qquad \zeta = \frac{s_y^2 - k s_x^2}{2 s_{xy}}$

maakt (9) minimaal, de waarde $\zeta - \sqrt{\zeta^2 + k}$ maakt (9) maximaal.

Het aantal graden van vrijheid (d.w.z. het aantal onafhankelijke quadraten) van (9) is n-2; als schatting voor δ^2 neemt men daarom de minimale quadraatsom (9) gedeeld door n-2, dus (na enige herleiding)

(14) $$\delta^{*2} = \frac{n}{n-2} \frac{s_y^2 - a^* s_{xy}}{k} = \frac{n}{n-2} \frac{a^* s_x^2 - s_{xy}}{a^*}$$

Opmerking: Indien er geen c is met $g_i = c h_i$ voor iedere i kan men gebruik maken van een benaderingsmethode; zie b.v. W.E. DEMING [4] en (voor hetzelfde probleem opgelost met de maximum likelihoodmethode) R.S. KOSHAL [9].

Bruikbaarheid van de schattingen; verzwakking van de voorwaarden.

4.2. Onder bovengenoemde onderstellingen verkrijgt men door toepassing van de methode der maximum-likelihood (vertaling: methode der aannemelijkste schattingen) dezelfde schattingen voor α en β [1]) (zie D.v.DANTZIG [3] p. 234 (Whr 323) e.v. en LINDLEY [12] p. 235 e.v.). Aangezien echter aan de op p. 221 (Whr 310) van D.v. DANTZIG (l.c.) vermelde voorwaarden hier niet voldaan is (het aantal onbekende parameters, waartoe immers ook de ξ_i behoren, gaat met $n \to \infty$ zelf naar oneindig), kan men hieruit niet, zoals vaak wel het geval is, tot bepaalde eigenschappen van genoemde schattingen besluiten. Deze moeten dus apart onderzocht worden en daarbij zal blijken, dat de voorwaarde van normaliteit, overbodig is, d.w.z. op geen enkele plaats gebruikt wordt voor de hier te bewijzen eigenschappen. (Zie Lindley l.c. pag. 237. A.A. MARKOFF [13] ontwikkelde reeds in 1910 de theorie der kleinste quadraten zonder normaliteitsonderstellingen).
Wij hebben nl. als de voorwaarde van normaliteit vervangen wordt door $\mathcal{E}\, u_i = 0$ voor iedere i en de voorwaarde van onafhankelijkheid van de meetfouten door ongecorreleerdheid:

$$\mathcal{E}\, s_x^2 = \frac{1}{n} \Sigma g_i\, \mathcal{E}(x_i - \bar{x})^2$$

---

[1]) De aannemelijkste schatting van δ^2 echter wijkt af van de kleinste-quadraten-schatting en is zelfs niet bruikbaar.

waarin (met $\bar{\xi} = \frac{1}{n}\Sigma g_i \xi_i$ en $\bar{\underline{u}} = \frac{1}{n}\Sigma g_i \underline{u}_i$):

$$\mathcal{E}(\underline{x}_i - \bar{\underline{x}})^2 = \mathcal{E}\{(\xi_i - \bar{\xi}) + (\underline{u}_i - \bar{\underline{u}})\}^2 = (\xi_i - \bar{\xi})^2 + \frac{n-1}{n}\frac{\delta^2}{g_i}$$

is dus

$$(15)\quad \begin{cases} \mathcal{E}\,\underline{s}_x^2 = s_\xi^2 + \frac{n-1}{n}\delta^2 \\ \mathcal{E}\,\underline{s}_y^2 = s_\eta^2 + \frac{n-1}{n}k\delta^2 \\ \mathcal{E}\,\underline{s}_{xy} = s_{\xi\eta} \end{cases}$$

waarin

$$s_\xi^2 = \frac{1}{n}\Sigma g_i(\xi_i - \bar{\xi})^2 \qquad s_\eta^2 = \frac{1}{n}\Sigma g_i(\eta_i - \bar{\eta})^2$$

en $$s_{\xi\eta} = \frac{1}{n}\Sigma g_i(\xi_i - \bar{\xi})(\eta_i - \bar{\eta})$$

is. De laatste vergelijkingen van (15) worden bewezen op analoge wijze als de eerste.

Verder volgt uit $\eta_i = \alpha\xi_i + \beta$ gemakkelijk

$$(16)\quad s_\eta^2 = \alpha s_{\xi\eta} = \alpha^2 s_\xi^2$$

Daar

$$\underline{s}_x^2 = \frac{1}{n}\Sigma g_i(\underline{x}_i - \bar{\underline{x}})^2 = \frac{1}{n}\Sigma g_i\{(\xi_i - \bar{\xi}) + (\underline{u}_i - \bar{\underline{u}})\}^2 =$$

$$= s_\xi^2 + \frac{2}{n}\Sigma g_i(\xi_i - \bar{\xi})(\underline{u}_i - \bar{\underline{u}}) + \underline{s}_u^2$$

is, convergeert $\underline{s}_x^2$ onder algemene voorwaarden stochastisch tot $\mathcal{E}\,\underline{s}_x^2$. Aangezien Lindley deze voorwaarden slechts vaag aanduidt, gaan we hier iets nader op in.

De verwachting van de tweede term is gelijk aan 0 en voor $n \to \infty$ gaat, zoals na enige herleiding blijkt, de spreiding van deze term naar 0 indien voor $n \to \infty$ $s_\xi^2 = O(n)$ en $\sum_1^n g_i = O(n^{3/2})$ is . Dan convergeert de term dus stochastisch naar 0. De laatste term , $\underline{s}_u^2$, blijkt stochastisch naar $\frac{n-1}{n}\delta^2$ te convergeren, indien voldaan is aan

$$\lim_{n\to\infty} \frac{1}{n^2}\Sigma g_i\,\mathcal{E}\,\underline{u}_i^4 = 0$$

daar in dat geval de spreiding van $\underline{s}_u^2$ voor $n \to \infty$ naar 0 convergeert en $\mathcal{E}\,\underline{s}_u^2 = \frac{n-1}{n}\delta^2$ is.

Onder analoge voorwaarden convergeert $\underline{s}_y^2$ stochastisch naar $\mathcal{E}\,\underline{s}_y^2$, terwijl het bestaan van de spreiding van $\underline{u}_i$ voor iedere i reeds voldoende is (tezamen met $\mathcal{E}\,\underline{u}_i = 0$ en onafhankelijkheid van de $\underline{u}_i$ en $\underline{v}_i$) voor stochastische convergentie van $\underline{s}_{xy}$ tot $s_{\xi\eta}$.

Is aan deze voorwaarden voldaan, dan volgt met behulp van (15) en (16) direct, dat de door (11), (13) en (14)

gegeven schattingen bruikbaar zijn.

Derhalve geldt:

Stelling 1: De met behulp van de methode der kleinste quadraten verkregen schattingen:

(13) $a^* = \zeta + \sqrt{\zeta^2 + k}$ met $\zeta = \dfrac{s_y^2 - k s_x^2}{2 s_{xy}}$

(11) $b^* = \bar{y} - a^* \bar{x}$

(14) $\sigma^{*2} = \dfrac{n}{n-2} \dfrac{s_y^2 - a^* s_{xy}}{k}$

zijn bruikbaar, indien voldaan is aan de volgende voorwaarden:

1) De meetfouten u_i $(i = 1,\ldots,n)$ zijn onafhankelijk verdeeld met spreidingen $\delta/\sqrt{g_i}$ (met bekende g_i) en gemiddelden 0.

2) De meetfouten v_i $(i = 1,\ldots,n)$ zijn onafhankelijk verdeeld met spreidingen $\delta\sqrt{k/g_i}$ (met bekende k) en gemiddelden 0.

3) u_i en v_j zijn ongecorreleerd voor iedere i en j.

4)
$$\lim_{n\to\infty} \frac{1}{n^2}\left(\sum_1^n g_i\right)^2 = \lim_{n\to\infty} \frac{1}{n} s_\xi^2 = \lim_{n\to\infty} \frac{1}{n^2} \sum_1^n g_i \mathcal{E} u_i^4 = \lim_{n\to\infty} \frac{1}{n^2} \sum_1^n g_i \mathcal{E} v_i^4$$

4.3. Bijzonder gevallen; foutloze ξ.

4.3.1. Indien ξ exact meetbaar is ($u_i \equiv 0$ voor iedere i) krijgt men een bijzonder geval van het vorige ($k\to\infty$; $k\delta^2 \to$ constante > 0). Noemen we in dit geval $\sigma_v^2 = \sigma^2$, dan vinden we uit (10) en (11) door $k\to\infty$ te laten gaan:

$$\left.\begin{array}{ll} (17) & a^* = \dfrac{s_{\xi y}}{s_\xi^2} \\ (18) & b^* = \bar{y} - a^* \bar{\xi} \end{array}\right\} \quad \text{met} \quad \left\{\begin{array}{l} \bar{y} = \frac{1}{n}\Sigma g_i y_i \\ \bar{\xi} = \frac{1}{n}\Sigma g_i \xi_i \\ s_\xi^2 = \frac{1}{n}\Sigma g_i (\xi_i - \bar{\xi})^2 \\ s_{\xi y} = \frac{1}{n}\Sigma g_i (\xi_i - \bar{\xi})(y_i - \bar{y}) \end{array}\right.$$

waarin $\xi_1,\ldots,\xi_n$ bekende parameters zijn.

Als schatting van σ^2 vinden we na enige herleiding van de vorm

$$\frac{1}{n-2}\Sigma g_i (y_i - a^*\xi_i - b^*)^2 \quad :$$

$$(19) \qquad s^{*2} = \frac{n}{n-2}(s_y^2 - a^* s_{\xi\eta})$$

In dit eenvoudiger geval zijn nog minder onderstellingen nodig, om de methode te rechtvaardigen. Nu gelden nl. de volgende stellingen:

Stelling 2: Indien voor iedere i geldt: $u_i \equiv 0$ en $\Sigma v_i = 0$ zijn de schattingen a^* en b^* uit (17) en (18) zuivere schattingen van α en β (bij iedere keuze der g_i).
Stelling 3: Indien voor iedere i geldt: $u_i \equiv 0$ en $\Sigma v_i = 0$ terwijl $\sigma_{v_i}^2 = \sigma^2/g_i$ is met bekende g_i (en onbekende σ), en de meetfouten v_i onderling onafhankelijk zijn, zijn de schattingen a^* en b^* uit (17) en (18) zuiver en bezitten zij van alle zuivere schattingen, die lineair in de y_i zijn de kleinste spreidingen (zij zijn dus de doeltreffendste schattingen van alle schattingen, die lineair in de y_i zijn). Tevens is

(20) $$\frac{n}{n-2}\left(\frac{S_y^2}{S_\xi^2} - a^{*2}\right)$$

een zuivere schatting van het spreidingskwadraat van a^*.

Het bewijs van stelling 2 is triviaal. Voor het bewijs van stelling 3 vergelijke men b.v. J. NEYMAN and F.N.DAVID [15], waar men ook een historisch overzicht over deze en algemenere stellingen uit de theorie der kleinste quadraten vindt.

4.3.2. H. CRAMÉR [2] p. 549 e.v. bewijst de volgende van R.A. FISHER afkomstige stelling:

Stelling 4: Is voor iedere i : $u_i \equiv 0$ en v_i normaal verdeeld met gemiddelde 0 en spreiding σ dan bezitten a^* en b^* normale verdelingen met gemiddelde α resp. β en $\sqrt{n}$ spreidingsquadraten $\sigma^2/n\,S_\xi^2$ resp. σ^2/n ; verder bezit $(n-2)\,S^{*2}/\sigma^2$ een χ^2-verdeling met n-2 graden van vrijheid, terwijl $S_\xi\sqrt{n}\cdot\frac{a^*-\alpha}{S^*}$ en $\sqrt{n}\cdot\frac{b^*-\beta+(a^*-\alpha)\bar{\xi}}{S^*}$ beide verdeeld zijn volgens een Student-verdeling met n-2 graden van vrijheid.

Hieruit kan men dus betrouwbaarheidsgrenzen bepalen voor α, voor β bij gegeven α (N.B.: a^* en b^* zijn niet onafhankelijk verdeeld) en een betrouwbaarheidsgebied voor α en β gezamenlijk. Tevens volgt uit deze stelling, dat voor dit geval ($g_i = 1$ voor iedere i) S^{*2} een zuivere schatting van σ^2 is.

4.3.3. Een ander bijzonder geval is k=1; men verkrijgt dan de orthogonale regressielijn. Voor dit geval zijn, evenals voor het algemene, geen betrouwbaarheidsgrenzen voor α en β bekend. De schattingen a^* en b^* bezitten de in punt

4.2 beschreven eigenschappen. Diverse andere methoden, waarbij een onderstelling over de spreidingen van de fouten in beide richtingen wordt gemaakt, zijn ontwikkeld. Een aantal publicaties dienaangaande vindt men bij A. WALD [17] p. 285 vermeld.

4.3.4. Uit het vorige blijkt, dat de methode der kleinste quadraten een uitstekende oplossing van het probleem geeft, als ξ exact meetbaar is, vooral als men bovendien normaliteit van de fouten in de η aan mag nemen. Voor het verkrijgen van een schatting van α en β is de methode ook zeer geschikt, indien ook ξ aan meetfouten onderhevig is, maar de verhouding van de spreidingen der meetfouten bekend is. Gegevens over de al of niet normaliteit van de fout in de η -richting en over de verhouding van de spreidingen kan men verkrijgen indien het aantal waargenomen punten niet te klein is en ieder punt P_i twee maal waargenomen wordt. De in het volgende te bespreken schattingsmethoden zijn daarom vooral van belang voor het geval, dat beide coördinaten met meetfouten belast zijn terwijl er geen of weinig duplicaatmetingen voorhanden zijn. De methoden ter verkrijging van betrouwbaarheidsintervallen, die besproken worden in de punten 8 en 9, zijn ook toepasbaar als de hier besproken methode der kleinste quadraten en die van Wald (zie punt 5) falen.

5. Methode van A. WALD.

5.1. A. WALD [17] was de eerste, die (in 1940) een schattingsmethode ontwierp, om een bruikbare schatting van de parameters te verkrijgen, als de verhouding van de spreidingen niet bekend is. De fouten behoeven ook niet normaal verdeeld te zijn; voor het geval ze dat wel zijn, leidt Wald bovendien betrouwbaarheidsintervallen af voor α , voor β bij gegeven α en een betrouwbaarheidsgebied voor α en β tezamen. Een kleine wijziging ter vergroting van de efficiency is later (1949) aangegeven door M.S. BARTLETT [1].

5.2. Wald maakt de volgende onderstellingen: Zij n even (voor oneven n kan men b.v. het waargenomen punt met de middelste x -waarde buiten beschouwing laten) en zij $m = \frac{1}{2}n$. Rangschik de punten P_i $(i = 1,\ldots,n)$ volgens opklimmende

x- (of y-) waarden en verdeel z e in twee groepen G_1 en G_2 waarvan G_1 de eerste m en G_2 de laatste m van deze gerangschikte punten bevat. Zij $P_1, \ldots, P_m$ de groep G_1 en $P_{m+1}, \ldots, P_n$ groep G_2. De onderstellingen zijn nu:

1) De meetfouten $\underline{u}_i$ (i = 1,...,n) hebben alle dezelfde verdeling met gemiddelde 0 en (onbekende) spreiding $\sigma_{\underline{u}}$ en zijn onderling ongecorreleerd, d.w.z. $\mathcal{E}\, \underline{u}_i \underline{u}_j = 0$ voor $i \neq j$. Analoog voor de $\underline{v}_i$ (met spreiding $\sigma_{\underline{v}}$).

2) $\mathcal{E}\, \underline{u}_i \underline{v}_j = 0$ voor iedere i en j.

3) $\liminf_{m \to \infty} \left| \frac{1}{m} \left\{ \sum_1^m \xi_i - \sum_{m+1}^n \xi_i \right\} \right| > 0$

4) Behoort P_i tot G_1 en P_j tot G_2, dan is $\xi_i < \xi_j$.

Een korte discussie van deze onderstellingen volgt in punt 5.5.

5.3. Als schattingen voert Wald nu in:

voor α:

$$(21) \qquad a_w = \frac{\sum_1^m y_i - \sum_{m+1}^n y_i}{\sum_1^m x_i - \sum_{m+1}^n x_i} \qquad m = \tfrac{1}{2} n$$

voor β:

$$(22) \qquad b_w = \bar{y} - a_w \bar{x} \quad \text{met } \bar{y} = \tfrac{1}{n} \Sigma y_i \text{ en } \bar{x} = \tfrac{1}{n} \Sigma x_i$$

voor $\sigma_{\underline{u}}^2$:

$$(23) \qquad S_{\underline{u},w}^2 = \frac{n}{n-1} \left\{ S_x^2 - \frac{S_{xy}}{a_w} \right\}$$

met $\begin{cases} S_x^2 = \frac{1}{n} \Sigma (x_i - \bar{x})^2 \\ S_y^2 = \frac{1}{n} \Sigma (y_i - \bar{y})^2 \\ S_{xy} = \frac{1}{n} \Sigma (x_i - \bar{x})(y_i - \bar{y}) \end{cases}$

en voor $\sigma_{\underline{v}}^2$:

$$(24) \qquad S_{\underline{v},w}^2 = \frac{n}{n-1} \left\{ S_y^2 - a_w S_{xy} \right\}$$

De schattingen $\underline{a}_w$ en $\underline{b}_w$ zijn, zoals Wald aantoont, bruikbare schattingen van α en β. Indien $\underline{S}_x^2$, $\underline{S}_y^2$ en $\underline{S}_{xy}$ in waarschijnlijkheid naar hun verwachtingswaarden convergeren, waartoe een verdere voorwaarde, zoals b.v.

$$\lim_{n \to \infty} \frac{1}{n} s_{\xi}^2 = 0 \qquad \mathcal{E}\, \underline{u}_i^4 \text{ en } \mathcal{E}\, \underline{v}_i^4 \text{ zijn eindig}$$

voldoende is (vgl. voorwaarde 4 van stelling 1), dan zijn ook $\underline{S}_{\underline{u},w}^2$ en $\underline{S}_{\underline{v},w}^2$ bruikbare schattingen van $\sigma_{\underline{u}}^2$ en $\sigma_{\underline{v}}^2$. Een dergelijke voorwaarde wordt door Wald niet genoemd en uit zijn bewijzen blijkt niet, hoe zij gemist kan worden.

5.4. Voor het geval, dat u en v normaal verdeeld zijn, leidt Wald betrouwbaarheidsintervallen af voor α en β. Dit levert geen nieuwe gezichtspunten op; voor de formules zij verwezen naar het geciteerde artikel.

5.5. Discussie van de voorwaarden; aanvulling van BARTLETT.

De schatting a_w is de helling van de lijn, die de "zwaartepunten" van G_1 en G_2 verbindt (als men aan alle punten een gelijk gewicht toekent). BARTLETT [1] stelt voor, de punten in 3 gelijke groepen te verdelen, H_1, H_2 en H_3, waarvan H_1 de punten met kleinste abscissen en H_3 de punten met grootste abscissen bevat.[1]) Hij bewijst voor een speciaal geval, dat dit de doeltreffendheid van de methode ten goede komt, d.w.z. dat de spreiding van de schatting erdoor vermindert. Verder geeft hij een duscussie van de mogelijkheden voor de schatting van σ_u en σ_v en leidt hij de formules voor de betrouwbaarheidsintervallen af, waartoe deze methode, naar analogie van die van Wald, leidt.

Voorwaarde 3) van Wald gaat bij Bartlett over in

$$\lim_{n\to\infty} \inf \left| \frac{1}{n} \left\{ \sum_{1}^{\frac{1}{3}n} \xi_i - \sum_{\frac{2}{3}n+1}^{n} \xi_i \right\} \right| > 0$$

en voorwaarde 4) in:

behoort P_i tot H_1, P_j tot H_2 en P_k tot H_3, dan is

$$\xi_i < \xi_j < \xi_k$$

Voorwaarde 3) kan, indien ξ, zoals ondersteld, geen waarschijnlijkheidsverdeling bezit, geen moeilijkheden opleveren. Zij houdt, daar n uiteraard eindig blijft bij toepassingen, in, dat men de punten niet te veel op een kluitje moet nemen.

Voorwaarde 4) is bezwaarlijker. Deze houdt in, dat de verdeling in de groepen G_1 en G_2, resp. H_1, H_2 en H_3, onafhankelijk van de meetfouten moet zijn of, populair gezegd, dat geen punt ten gevolge van de meetfout in de ξ-richting in een "verkeerde" groep mag kunnen raken. Deze voorwaarde is nodig, om de onafhankelijkheid van de meetfouten ondanks de groepering te behouden.

---

[1]) Deze suggestie werd oorspronkelijk op experimentele gronden in iets gewijzigde vorm door K.R. NAIR en K.S. BANERJEE [15], geopperd.

Aan deze voorwaarde is is b.v. bij een normaal verdeelde fout niet voldaan. Het is echter duidelijk, dat ook, als men weet, dat er slechts een zeer geringe kans q bestaat op deze verwisseling, de methode bruikbaar blijft; de onbetrouwbaarheidsdrempel van de betrouw baarheidsintervallen neemt dan ten hoogste met q toe. Wald geeft van dit punt een vrij uitvoerige discussie (l.c. pag. 294 e.v.), die hier niet gereproduceerd zal worden. Als kleine wijziging op de methode van Bartlett zou men, als q niet voldoende klein is en men wel beschikt over een bovengrens voor de "range" (variatiebreedte) van u, in sommige gevallen als volgt te werk kunnen gaan: verdeel de punten in de volgende groepen:

K_i, bestaande uit die punten, die zeker tot H_i behoren ($i = 1,2,3$) [1];
K_4, bestaande uit punten, die tot H_1 of H_2 behoren;
K_5, bestaande uit punten, die tot H_2 of H_3 behoren.

Kies nu stochastisch ("at random") een zo groot aantal punten uit K_4 als nodig is om K_1 $\frac{1}{3}n$ punten te geven en voeg deze bij K_1 en analoog uit K_5 om K_3 te completeren. Gebruiken wij nu K_1 en K_3 in plaats van H_1 en H_3 dan behouden wij (gedeeltelijk) het voordeel van de grotere efficiency van Bartlett's methode en maken tevens $q = 0$. Hierbij is ondersteld, dat u begrensd is; is dit niet zo, dan kan men toch veelal op deze w ijze de q voldoende klein maken.

Uit het vorige blijkt, dat de methode van Wald het best toepasbaar is, indien de "puntenwolk" $P_1,\ldots,P_n$ de vorm van een halter heeft, twee puntenbollen met een gelijk aantal punten op zo groot mogelijke afstand. Soms zal men hiermee bij de inrichting van een experiment rekening kunnen houden. Voor de bepaling van q is het bovendien zeer gewenst alle waarnemingen in duplo te verrichten.

6. Methode van G.W. HOUSNER en J.F. BRENNAN.

5.1. G.W. HOUSNER en J.F. BRENNAN [8] ontwikkelden een schattingsmethode, verwant aan die van Wald. Zij behandelden verder

---

[1]) Strikt genomen, dient men toe te voegen: en die ook niet, ten gevolge van de meetfout zo hadden kunnen uitvallen, dat men hierover geen zekerheid bezit.

een speciaal voorbeeld, waarbij de doeltreffendheid van hun schatting groter is dan die van Wald. Een algemene discussie van deze efficiency geven zij niet.

6.2. Onderstellingen:

1) Onderstellingen 1) en 2) van Wald (zie 5.2).
2) Voor $i<j$ is $\xi_i < \xi_j$ [1]).
3) Er is een $\varepsilon>0$ zo, dat

$$\lim_{n\to\infty} P\left[\tfrac{1}{n^2}\left|\Sigma\, i(x_i-\bar{x})\right| > \varepsilon\right] = 1 \quad \text{a.s.}$$

(Deze derde onderstelling wordt door hen als vanzelfsprekend aangenomen).

6.3. De schattingen:

Voor α:

$$(25) \qquad a_{HB} = \frac{\sum\limits_{k<i}\sum (y_i - y_k)}{\sum\limits_{k<i}\sum (x_i - x_k)} = \frac{\sum\limits_{i=1}^{n} i(y_i-\bar{y})}{\sum\limits_{i=1}^{n} i(x_i-\bar{x})}$$

Voor β:

$$(26) \qquad b_{HB} = \bar{y} - a_{HB}\,\bar{x}$$

Van beide schattingen bewijzen zij de bruikbaarheid.

Housner en Brennan geven de formules in iets algemenere vorm, nl. voor het geval, dat er van het punt $P_i \equiv (\xi_i, \eta_i)$ n_i waarnemingen zijn verricht ($i = 1,\ldots,n$).

6.4. Opmerkingen: Voorwaarde 2) komt overeen met voorwaarde 4) van Wald, maar eist meer. Voorwaarde 3) is het analogon van voorwaarde 4) van Wald. Deze methode zal in verband met voorwaarde 2) het best toepasbaar zijn, als de punten ongeveer aequidistant liggen, daar dan de kans op volgorde-verwisselingen (zo klein mogelijk is, dit in tegenstelling) tot de methode van Wald, waarbij verwisselingen binnen ieder van de twee groepen zijn toegestaan.

---

[1]) D.w.z. de punten zijn gerangschikt volgens opklimmende waarden van de "ware" abcsissen. I.h.a. zal men tot deze rangschikking alleen in staat zijn door de punten volgens opklimmende x-waarden te ordenen; voorwaarde 2) houdt dan in, dat de verschillen tussen de ξ -waarden zo groot zijn, dat de meetfouten de volgorde niet kunnen verstoren.

7. Onzuiverheid van de schattingen $\underline{a}^*$, $\underline{a}_W$ en $\underline{a}_{HB}$:

7.1. Van de kleinste-quadraten-schatting (in het algemene geval), en de schattingen van Wald en Housner en Brennan wordt slechts de bruikbaarheid door de auteurs bewezen, aan de zuiverheid of onzuiverheid wordt geen aandacht geschonken. Men kan hierover echter het volgende opmerken:

Verwisselt men bij genoemde methoden de twee coördinaatassen en past men vervolgens dezelfde schattingsmethode toe, dan verkrijgt men dezelfde lijn als schatting van I als oorspronkelijk. [1] Met andere woorden, men vindt één lijn en niet twee verschillende, zoals in d. gewone regressieanalyse. Dit feit leidt er echter toe, dat de schatting in het algemeen noodzakelijkerwijze onzuiver is, zoals men als volgt kan inzien:

De vergelijking:

(1) $$\eta = \alpha \xi + \beta$$

kan men ook schrijven als

(27) $$\xi = \frac{1}{\alpha}\eta - \frac{\beta}{\alpha}$$

Is nu $\underline{a}$ de schatting van α, dan vindt men door verwisseling van de assen $\frac{1}{\underline{a}}$ als schatting van $\frac{1}{\alpha}$. De eigenschappen, die aan $\underline{a}$ ten opzichte van α toekomen, moeten dus ook aan $\frac{1}{\underline{a}}$ ten opzichte van $\frac{1}{\alpha}$ toekomen, dus als

$$\mathcal{E}\,\underline{a} = \alpha$$

is, dan zou ook

$$\mathcal{E}\,\frac{1}{\underline{a}} = \frac{1}{\alpha}$$

moeten zijn. Daar echter (indien $P[\underline{a} > 0] = 1$ is) geldt:

(28) $$\mathcal{E}\,\frac{1}{\underline{a}} > \frac{1}{\mathcal{E}\,\underline{a}}$$

tenzij $\underline{a}$ één waarde met waarschijnlijkheid 1 aanneemt, is dit in het algemeen niet het geval.

7.2. In plaats van de gewone zuiverheidseigenschap:

$$\mathcal{E}\,\underline{a} = \alpha$$

kan men echter ook de eigenschap van "mediaan-zuiverheid"

---

[1]) Bij de methoden van Wald en Housner en Brennan geldt dit slechts, als de groepering resp. volgorde bij rangschikking naar opklimmende x-waarden dezelfde is als bij rangschikking naar opklimmende y-waarden.

beschouwen; wij kunnen $\underline{a}$ mediaan-zuiver noemen, als geldt:

(29) $\quad Med\ \underline{a} = \alpha$

d.w.z. als de mediaan van de verdeling van $\underline{a}$ gelijk aan α is. Daar voor de mediaan (mits $P[\underline{a}>0]=1$ of 0 is) geldt:

(30) $\quad Med\ \frac{1}{\underline{a}} = \frac{1}{Med\ \underline{a}}$

kan de eigenschap van mediaan-zuiverheid bij schattingen van de in het vorige punt beschreven aard wel vervuld zijn.

Inderdaad geldt voor de schattingen $\underline{a}_W$ en $\underline{a}_{HB}$ van Wald en van Housner en Brennan:

(31) $\quad Med\ \underline{a}_W = Med\ \underline{a}_{HB} = \alpha$

Immers in beide gevallen is:

$$\underline{y}_i = \alpha \underline{x}_i - \beta - (\alpha \underline{u}_i - \underline{v}_i)$$

dus

$$\underline{y}_i - \underline{y}_k = \alpha(\underline{x}_i - \underline{x}_k) - \alpha(\underline{u}_i - \underline{u}_k) + (\underline{v}_i - \underline{v}_k)$$

dus:

(32) $$\underline{a}_W = \alpha - \frac{\alpha\left(\sum_1^m \underline{u}_i - \sum_{m+1}^n \underline{u}_i\right) + \sum_1^m \underline{v}_i - \sum_{m+1}^n \underline{v}_i}{\sum_1^m \underline{x}_i - \sum_{m+1}^n \underline{x}_i}$$

en

(33) $$\underline{a}_{HB} = \alpha - \frac{\alpha \sum\sum_{k<i} (\underline{u}_i - \underline{u}_k) + \sum\sum_{k<i} (\underline{v}_i - \underline{v}_k)}{\sum\sum_{k<i} (\underline{x}_i - \underline{x}_k)}$$

In beide vergelijkingen is de teller van de tweede term van het rechterlid symmetrisch verdeeld (daar $\underline{u}_i - \underline{u}_k$ dit), is, als $\underline{u}_i$ en $\underline{u}_k$ dezelfde verdeling bezitten. Daar de verdeling van de noemer niet onafhankelijk is van die van de teller, kan men niet concluderen, dat deze breuken symmetrisch verdeeld zijn, maar wel, dat hun mediaan gelijk aan nul is, daar de noemers in beide gevallen steeds positief zijn en de tellers de mediaan 0 hebben.

7.3. In beide gevallen is

(34) $\quad \underline{b} = \beta + (\alpha - \underline{a})\bar{\xi} - (\underline{a}\bar{\underline{u}} - \bar{\underline{v}})$

Daar echter de mediaan niet, zoals het gemiddelde, de eigenschap van additiviteit bezit en $\alpha - \underline{a}$ in het algemeen niet symmetrisch verdeeld is (bij symmetrisch verdelingen is de mediaan wel additief), lijkt het niet eenvoudig, voorwaarden

aan te geven, die $\underline{b}$ tot een mediaan-zuivere schatting van β maken.

8. Betrouwbaarheidsgrenzen voor α en β zonder onderstelling van normaliteit van de fouten.

8.1. Op de Statistische Afdeling van het Mathematisch Centrum zijn enige methoden ontworpen ter berekening van betrouwbaarheidsgrenzen voor α en β zonder de voorwaarde van normaliteit van de meetfouten, die in dit en het volgende punt besproken worden.

8.2. De onderstellingen, die bij de eerste methode (J. HEMELRIJK [7]) gebruikt worden, zijn verschillend bij het bepalen van betrouwbaarheidsgrenzen voor α en voor β .

Voor het bepalen van een betrouwbaarheidsgebied voor α wordt ondersteld:

1a) De meetfouten u_i en v_i hebben een simultane waarschijnlijkheidsverdeling, die onafhankelijk is van i.[1])

1b) De kans, dat het punt (u_i, v_i) op een gegeven rechte lijn in het (u_i, v_i)-vlak ligt, is gelijk aan 0 voor iedere lijn in dit vlak.

1c) Noemen wij het "ware" punt met kleinste abscis Q_1 en het "ware" punt met grootste abscis Q_n, dan kan men de bij Q_1 en Q_n behorende punten P_1 en P_n aanwijzen onder de punten P_i $(i = 1,\ldots,n)$.[2])

Voor het bepalen van een betrouwbaarheidsinterval voor β onder de hypothese, dat de helling α gelijk is aan een gegeven waarde a wordt ondersteld:

2) De meetfouten u_i en v_i hebben voor iedere i een simultane waarschijnlijkheidsverdeling, die van i afhankelijk kan zijn, maar waarbij voor iedere i de kans, dat P_i aan de éne zijde van L ligt gelijk is aan de kans, dat P_i aan de andere

---

[1]) u_i en v_i behoeven dus bij een zelfde i niet onafhankelijk van elkaar verdeeld te zijn; wel is de verdeling van de paren (u_i, v_i) en (u_j, v_j) voor $j \neq i$ onafhankelijk.

[2]) Deze onderstelling is van dezelfde aard als onderstelling 4) van Wald (zie punt 5.2), zij het iets zwakker. Zij kan nog enigszins worden verzwakt; dit zal hier echter niet worden behandeld. In de genoemde publicatie wordt hierover iets meer gezegd.

zijde ligt, terwijl de kans, dat P_i op L ligt gelijk is aan 0.

Voor de bepaling van een betrouwbaarheidsgebied voor α en gezamenlijk wordt zowel 1a)..,c) als 2) ondersteld.

8.3. Betrouwbaarheidsgebied voor α.

Een betrouwbaarheidsgebied A voor α met onbetrouwbaarheidsdrempel

(35) $$p_1 = \frac{(m+1)(m+2)}{n(n-1)} \qquad (0 \leq m \leq n-2)$$

wordt nu gevormd door die waarden a, die voldoen aan het volgende criterium:

Trekt men in het (ξ, η)-vlak(dat tevens (x,y)-vlak is), door P_1 en P_n twee evenwijdige lijnen L_1 en L_n met richti coëfficient a, dan bevat de strook, die begrensd wordt do L_1 en L_n (deze lijnen niet inbegrepen) hoogstens n-m-3 van de punten P_i (i = 2,..,n-1).

Bewezen dient dan te worden dat

(36) $$P[\alpha \in A] = \frac{(m+1)(m+2)}{n(n-1)}$$

is. Dit bewijs berust op het feit, dat de afstand η_i van P_i tot L (i = 1,...,n), gemeten in een willekeurige vaste richting, onafhankelijk van elkaar verdeeld zijn met dezelfde verdelingsfunctie voor iedere i. Uit het feit, dat alleen dan α niet tot A behoort, als minstens n-m van de afstanden η_i gelegen zijn in het afgesloten interval met η_1 en η_n als grenzen, volgt (36) met behulp van een eenvoudige redenering. A bestaat uit een eindig aantal intervallen. Indien men door P_1 en P_n twee evenwijdige lijnen kan trekken, waartussen alle overige punten P_i gelegen zijn, gaat A over in één interval. Dit is steeds het geval als de meetfouten niet te groot zijn in vergelijking met de afstand van Q_1 en Q_n tot de overige punten Q_i.

8.4. Betrouwbaarheidsinterval voor β onder de hypothese $\alpha = a$.

Een betrouwbaarheidsinterval B voor β bij gegeven $\alpha = a$ met onbetrouwbaarheidsdrempel

(37) $$p_2 = 2^{-n+1} \sum_{i=0}^{k} \binom{n}{i} \qquad (0 \leq k < \frac{n-3}{2})$$

verkrijgt men door de lijnen L_1 en L_n door P_1 resp. P_n met

richtingscoëfficient a evenwijdig te verschuiven tot aan iedere zijde van de strook, die zij begrenzen, precies k van de punten P_i gelegen zijn, terwijl op beide nieuwe lijnen L_1 en L_n minstens één van de punten P_i ligt (zodat zij niet dichter bij elkaar gebracht kunnen worden, zonder dat aan minstens één van beide zijden het aantal buiten de strook gelegen punten groter dan k wordt). De doorsnijding van deze strook met de η -as is het betrouwbaarheidsinterval B. Het bewijs van deze stelling volgt gemakkelijk uit onderstelling 2).

8.5. Betrouwbaarheidsgebied voor α en β gezamenlijk.

De bepaling van het betrouwbaarheidsgebied voor α komt overeen met het vaststellen vaneen criterium ter verwerping van de hypothese, dat α de waarde a bezit, waarbij de kans p_1 bestaat, dat α zelf aan dit criterium voldoet. Evenzo komt de bepaling van het betrouwbaarheidsinterval voor β onder de voorwaarde $\alpha = a$ neer op het vaststellen van een criterium ter verwerping van de hypothese, dat β onder de voorwaarde $\alpha = a$ de waarde b bezit, waarbij de kans p_2 bestaat, dat β zelf aan dit criterium voldoet.

Van deze twee criteria is het eerste gebaseerd op de plaats, die z_1 en z_n innemen in de naar opklimmende grootte gerangschikte rij getallen ($i = 1,\ldots,n$) (het is dus onafhankelijk van het teken van de z_i), terwijl het tweede criterium invariant is tegen permutatie van de punten P_i (en juist van de tekens van de z_i afhangt). Hieruit volgt, dat de twee criteria onafhankelijk zijn, zodat de kans, dat hetzij α aan het eerste, hetzij β aan het tweede criterium (als daarbij van de ware richting α gebruik gemaakt wordt) voldoet, gelijk is aan

$$p = p_1 + p_2 - p_1 p_2 \tag{38}$$

Laat men derhalve a alle waarden van A doorlopen en bepaalt men bij iedere a het bijbehorende interval B dan vormt de zo verkregen verzameling van punten (a,b) een betrouwbaarheidsgebied door het punt (α,β) met onbetrouwbaarheidsdrempel p.

8.6. Ter verduidelijking van de methode geven wij in fig. 1 voor een eenvoudig geval de constructie weer.

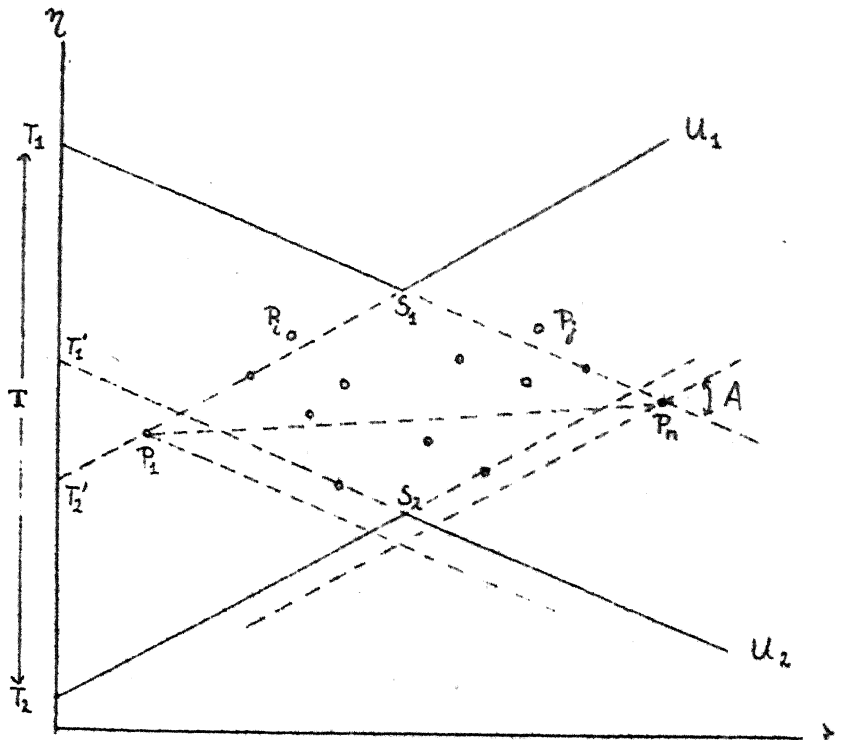


fig. 1 $n = 13$ $m = 1$ $\ell = 1$
$p_1 = 0{,}039$ $p_2 = 0{,}004$

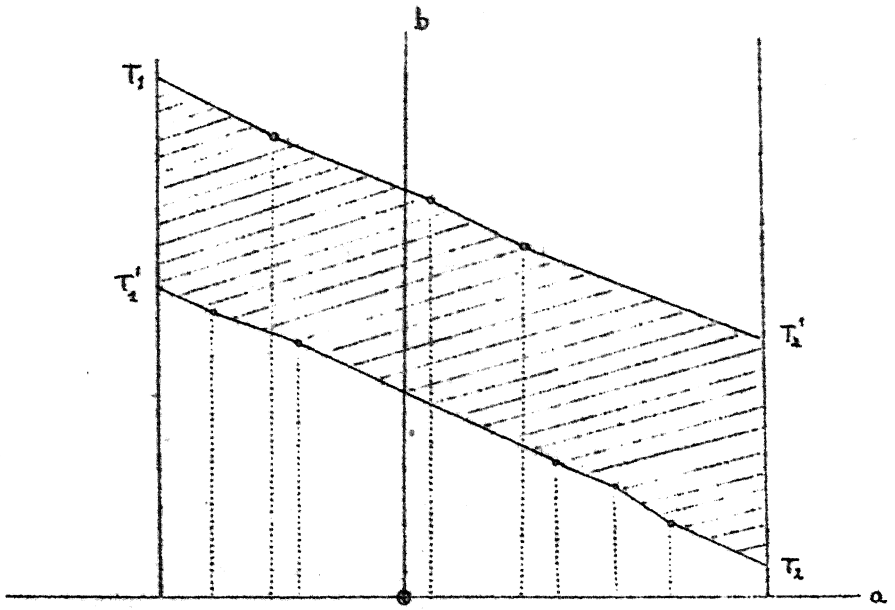
In deze figuur stelt A de hoek van richtingen voor, die een betrouwbaarheidsinterval met onbetrouwbaarheidsdrempel $p_1 = 0{,}039$ voor de richting van L vormen. Voor de beide uiterste richtingen zijn de intervallen (T_1, T_1') resp. (T_2, T_2') betrouwbaarheidsintervallen voor β onder de voorwaarde, dat α de bijbehorende waarde bezit, met onbetrouwbaarheidsdrempel $p_2 = 0{,}004$.

Het door de gebroken lijnen $T_1 S_1 u_1$ en $T_2 S_2 u_2$ begrensde gebied G van het (ξ, η)-vlak vormt nu een betrouwbaarheidsgebied voor L met onbetrouwbaarheidsdrempel $p_1 + p_2 - p_1 p_2 = 0{,}$(in die zin, dat de kans, dat dit stochastisch gebied L geheel bevat = 0,957 is, terwijl omgekeerd niet iedere lijn die er geheel in ligt, een a en b bezit, die tot het simultane betrouwbaarheidsgebied voor α en β behoren. Wel is dit met de a steeds het geval, maar de b behoeft niet tot het met behulp van deze a geconstrueerde betrouwbaarheidsinterval R te behoren. Het betrouwbaarheidsgebied voor L,

de verzameling van alle lijnen, die geheel in G liggen.

Een kleine complicatie doet zich verder voor, als een lijn als b.v. $P_i P_j$ in figuur 1 de lijnen $T_1 S_2$ en $U_1 S_2$ snijdt. Het van de hoek $T_1 S_2 U_1$ afgesneden gedeelte behoort dan ook bij G.

Het betrouwbaarheidsgebied voor α en β gezamenlijk kr in dit geval ongeveer de in figuur 2 getekende vorm (het gearceerde gedeelte van het (a,b)-vlak).



figuur 2.

8.7. De hier beschreven methode is, zoals uit de constructie en uit voorwaarde 1.c van 8.2 blijkt, vooral geschikt, indien de meeste punten vrij dicht bij elkaar liggen, terwijl zowel links als rechts van deze puntenwolk minstens één punt gelegen is. Dit is dus juist in die gevallen, waarvoor de methoden van Wald en die van Housner en Brennan, evenals de in het volgende punt nog te bespreken methoden minder doelmatig zijn (i.h.b. ook in verband met de voor de verschillende methoden variërende eisen omtrent invariantie van de volgorde t.o.v. de meetfouten).

Lineaire transformaties van het assenstelsel laten de wijze van constructie invariant. Dit kan het graphisch uitvoeren van de constructie soms vergemakkelijken. Nadat deze is uitgevoerd leiden dan enkele berekeningen van hellingen

tot de precieze betrouwbaarheidsintervallen.

Uit deze methode kunnen enkele eenvoudige toetsen voor bepaalde hypothesen omtrent L worden afgeleid.

9. Methoden van H. THEIL.

9.1. H. THEIL [16] heeft twee methoden ontworpen, die weliswaar opgezet zijn in het kader der regressieanalyse, maar die ook op het hier besproken probleem kunnen worden toegepast. Wij lichten uit zijn (nog niet verschenen) artikel die [illegible]len, die relevant zijn voor het hier gestelde probleem.

9.2. De eerste methode berust op de volgende onders

Voor de bepaling van een betrouwbaarheidsinterval

1) Is $\xi_i < \xi_j$ dan is ook $x_i < x_j$. (vgl. 6.2).

2) Eén van de beide volgende onderstellingen moet vervuld zijn: ')

a. De meetfouten $\underline{u}_i$ en $\underline{v}_i$ (i = 1,...,n) hebben voor iedere i dezelfde simultane waarschijnlijkheidsverdeling; de paren $(\underline{u}_i, \underline{v}_i)$ en $(\underline{u}_j, \underline{v}_j)$ zijn onafhankelijk voor $i \neq j$; de verdelingsfunctie van $\underline{v}_i - \alpha \underline{u}_i$ is continu in de mediaan.

b. De meetfouten zijn alle onafhankelijk van elkaar verdeeld; ieder van hen is bovendien symmetrisch verdeeld, terwijl de mediaan van $\underline{u}_i$ voor iedere i dezelfde is; evenzo de mediaan van $\underline{v}_i$ voor iedere i. De verdelingsfuncties zijn alle continu in hun medianen.

Voor de bepaling van een betrouwbaarheidsinterval voor β onder de hypothese $\alpha = a$, wordt bovendien ondersteld:

3) Med $(\underline{v}_i - \alpha \underline{u}_i) = 0$ voor iedere i.

9.3. De betrouwbaarheidsintervallen worden nu als volgt bepaald:

Zij de rangschikking $P_1, \ldots, P_n$ verricht naar opklimmende grootte van x_i en zij $m = \frac{1}{2}n$ (met eventuele weglating) van het middelste punt). Zij verder:

$$\underline{a}_i = \frac{\underline{y}_i - \underline{y}_{m+i}}{\underline{x}_i - \underline{x}_{m+i}} \tag{39}$$

Wegens (vgl. 7.2)

$$\underline{y}_i - \underline{y}_{m+i} = \alpha(\underline{x}_i - \underline{x}_{m+i}) - \alpha(\underline{u}_i - \underline{u}_{m+i}) + (\underline{v}_i - \underline{v}_{m+i})$$

') Dit is voldoende, maar noodzakelijk is strikt genomen slechts, dat $\alpha(\underline{u}_i - \underline{u}_{m+i}) - (\underline{v}_i - \underline{v}_{m+i})$ voor iedere i als mediaan 0 heeft.

is dan

(40) $\text{Med } \underline{a}_i = \alpha$

indien voorwaarde 1) en 2) vervuld zijn.

Rangschikt men nu de gevonden a_i (i = 1,...,n) volgens opklimmende waarden en is:

(41) $$p_1 = 2^{-m+1} \sum_{j=0}^{r-1} \binom{m}{j}$$

dan zijn de r^{de} en de $(m-r+1)^{de}$ van deze rij der a_i de uiteinden van een betrouwbaarheidsinterval voor α met onbetrouwbaarheidsdrempel p_1.

Voor β kan men, onder hypothese $\alpha = a$, een betrouwbaarheidsinterval afleiden op dezelfde wijze als bij de vorige methode (zie 8.4.). Bezit dit betrouwbaarheidsinterval de onbetrouwbaarheidsdrempel p_2 en duiden wij de twee intervallen aan met (a_1, a_2) en (b_1, b_2), dan is:

(42) $$P[\, \underline{a}_1 \leqq \alpha \leqq \underline{a}_2 \,;\, \underline{b}_1 \leqq \beta \leqq \underline{b}_2 \,] \geqq (1-p_1)(1-p_2)$$

zodat men een rechthoekig betrouwbaarheidsgebied voor α en β gezamenlijk heeft. De reden, dat men in dit geval geen betrouwbaarheidsgebied voor het punt (α,β) van de in 8.6 beschreven vorm verkrijgt, is dat in dit geval de in 8.5 genoemde onafhankelijkheidsvoorwaarde niet vervuld is.

9.4. De <u>tweede methode</u> bezit vermoedelijk het voordeel, scherper te zijn dan de vorige. De berekeningen worden echter veel uitgebreider, daar hierbij niet slechts van $\frac{1}{2}n$ hellingen gebruik wordt gemaakt, maar van de hellingen van alle mogelijke verbindingslijnen van puntenparen $P_i P_j$ (het verschil met de eerste methode, vertoont een zekere analogie met het verschil tussen de methode van Housner en Brennan en die van Wald).

Voor het bepalen van een <u>betrouwbaarheidsinterval voor α</u> worden de volgende onderstellingen gemaakt:

1) Is $\xi_i < \xi_j$ dan is $x_i < x_j$.

2) De meetfouten $\underline{u}_i$ en $\underline{v}_i$ bezitten voor iedere i dezelfde simultane verdeling; de paren $(\underline{u}_i, \underline{v}_i)$ en $(\underline{u}_j, \underline{v}_j)$ zijn onafhankelijk verdeeld voor $i \neq j$; de grootheid $\underline{v}_i - a\,\underline{u}_i$ is continu verdeeld voor iedere constante a (d.w.z. in het (u_i, v_i)-vlak is voor iedere vaste rechte lijn de kans, dat het punt $(\underline{u}_i, \underline{v}_i)$ er op ligt gelijk aan nul).

9.5. Het betrouwbaarheidsinterval (voor α) wordt nu als volgt bepaald:

De rangschikking $P_1, \ldots, P_n$ zij weer zodanig, dat $x_i < x_j$ is voor $i < j$. Zij verder

$$(43) \qquad \underline{a}_{ij} = \frac{\underline{y}_i - \underline{y}_j}{x_i - x_j} = \alpha + \frac{(\underline{v}_i - \underline{v}_j) - \alpha(\underline{u}_i - \underline{u}_j)}{x_i - x_j} \qquad i < j$$

met i = 1,..., n-1 en j = i+1,...,n. Dan is $\underline{a}_{ij} > \alpha$ dan en slechts dan als $\underline{v}_i - \alpha \underline{u}_i < \underline{v}_j - \alpha \underline{u}_j$. Vullen wij voor α een willekeurig aangenomen getal a in, dan is het aantal gevallen, waarvoor $v_i - a u_i > v_j - a u_j$ (met $i < j$) is, voor ieder waarnemingsresultaat bepaald. Dit aantal, dat behalve van het waarnemingsresultaat ook van a afhangt, noemen wij $q^{(a)}$; $\underline{q}^{(a)}$ is een stochastische variabele op de collectie van alle mogelijke waarnemingsresultaten.

Nu is de verdeling van

$$(44) \qquad \underline{\tau} = 1 - {}^{2}\underline{q}^{(a)} \Big/ \binom{n}{2}$$

bekend, onder de hypothese dat de grootheid $v_i - a\,\underline{u}_i$ onafhankelijk van i verdeeld is (zie M.G. KENDALL [10], [11]; τ heet de rank correlation coefficient, hetgeen men zou kunnen vertalen met rangcorrelatie coëfficient). Aan deze hypothese is in ons geval voldaan als $a = \alpha$ is, terwijl bij toenemende afwijking van α de kans op grotere q toeneemt. Neemt men nu r zo, dat

$$(45) \qquad P[\underline{q}^{(a)} \geqq r - 1] = \tfrac{1}{2} p_1$$

is, en rangschikt men de a_{ij} volgens opklimmende waarden

$$(a_{ij})_1 < \ldots < (a_{ij})_{\binom{n}{2}}$$

dan zijn de r^{de} en de $\left(\binom{n}{2} - r + 1\right)^{\text{de}}$ van deze rij de uiteinden van een betrouwbaarheidsinterval voor α met onbetrouwbaarheidsdrempel p_1:

$$(46) \qquad P[(\underline{a}_{ij})_r \leqq \alpha \leqq (\underline{a}_{ij})_{\binom{n}{2}-r+1}] = 1 - p_1.$$

Voor β gaat men te werk als bij de eerste methode.

9.5. Deze methoden zijn, zoals in de betreffende publicaties zal blijken, voor velerlei generalisatie vatbaar: het aantal variabelen, waarvan η lineair afhangt kan willekeurig groot gemaakt worden; de linea[illegible]riteit van het verband kan getoetst worden tegen convexiteit; uitbreiding tot vergelijkingen van hogere graad is mogelijk; stelsels vergelijkingen met onbekende parameters kunnen op deze wijze worden behandeld.

In verband met voorwaarde 1) zijn beide methodes, evenals die van Housner en Brennan, het best toepasbaar, als de punten P_i ongeveer aequidistant liggen.

10. Het geval, dat ξ een waarschijnlijkheidsverdeling bezit.

10.1. In 2.2 zijn twee verschillende situaties genoemd, aangeduid als geval I en geval II. Hoewel deze gevallen betrekking hebben op geheel verschillende problemen, kan er toch een nauw verband tussen gelegd worden.

In geval I bezit ξ geen waarschijnlijkheidsverdeling en worden de coördinaten van de punten $\alpha_1, \ldots, \alpha_n$ opgevat als onbekende parameters van de waarschijnlijkheidsverdeling van de $\underline{x}_i$ en $\underline{y}_i$. Dit doet zich b.v. voor als men door de inrichting van het experiment de grootte der ξ_i min of meer naar willekeur kan bepalen ook al kan men ze niet exact meten.

In geval II bezit ξ wèl een waarschijnlijkheidsverdeling (en we schrijven dan $\underline{X}$ in plaats van ξ en $\underline{Y}$ in plaats van η). Deze situatie kan zich b.v. voordoen, indien $\underline{X}$ en $\underline{Y}$ in principe meetbare grootheden zijn, maar men, om één of andere reden, alleen over onnauwkeurige metingen beschikt (de nauwkeurige meting kan b.v. een niet beschikbare apparatuur eisen of te veel tijd in beslag nemen). Daar $\underline{x}$ en $\underline{y}$ nu ook een simultane waarschijnlijkheidsverdeling bezitten, kan men de $\underline{X}$ en $\underline{Y}$ geheel uit het vraagstuk elimineren en aldus tot de gewone regressieanalyse overgaan. Indien echter $\underline{X}$ en $\underline{Y}$ b.v. natuurkundige grootheden voorstellen, is het zeer wel mogelijk, dat het verband tussen $\underline{X}$ en $\underline{Y}$ wel, maar dat tussen $\underline{x}$ en $\underline{y}$ niet van uiteindelijk belang is. Bovendien behoeft er tussen $\underline{x}$ en $\underline{y}$ geen lineair verband te bestaan, als dit met $\underline{X}$ en $\underline{Y}$ wel het geval is. LINDLEY [12] bewijst dienaangaande de volgende stelling:

Stelling 5: Indien $\underline{X}$, $\underline{u}$ en $\underline{v}$ onderling onafhankelijk verdeelde stochastische variabelen zijn, met $\mathcal{E}\underline{u} = \mathcal{E}\underline{v} = 0$ en indien geldt:

$$\underline{Y} = \alpha\underline{X} + \beta \; ; \quad \underline{x} = \underline{X} + \underline{u} \; ; \quad \underline{y} = \underline{Y} + \underline{v}$$

dan is noodzakelijk en voldoende, opdat de regressie van $\underline{y}$ ten opzichte van $\underline{x}$ lineair zij, d.w.z. opdat

(47) $$\mathcal{E}_{\underline{X}=x}\,\underline{Y} = \alpha' x + \beta'$$

is voor iedere x , dat voor de karakteristieke functies

$$Z_{\underline{U}}(\tau) \overset{df}{=} \mathcal{E}\, e^{\tau \underline{U}} \quad \text{en} \quad Z_{\underline{\tilde{X}}}(\tau) \overset{df}{=} \mathcal{E}\, e^{\tau \underline{\tilde{X}}} \qquad ^{1})$$

van $\underline{U}$ resp. $\underline{\tilde{X}} = \underline{X} - \mathcal{E}\underline{X}$, geldt:

(48) $$Z_{\underline{\tilde{X}}}(\tau) = \left\{ Z_{\underline{U}}(\tau) \right\}^{\frac{\alpha'}{\alpha - \alpha'}}$$

en

(49) $$\beta' = \beta + (\alpha - \alpha').\, \mathcal{E}\underline{X}.$$

Opmerking: 1. De in deze stelling genoemde noodzakelijke en voldoende voorwaarde betekent b.v. voor het geval, dat $\underline{X}$ normaal verdeeld is, dat $\underline{U}$ ook een normale verdeling moet bezitten. Daar bovendien in het algemeen een macht van een karakteristieke functie zelf geen karakteristieke functie is, wordt door (48) aan het type van de verdeling van $\underline{X}$ een sterke beperking opgelegd.

2. Uit de stelling volgt verder direct, dat

$$\frac{\alpha'}{\alpha - \alpha'} > 0 \qquad \text{dus} \qquad 0 < |\alpha'| < |\alpha|$$

moet zijn, terwijl α' en α hetzelfde teken bezitten.

Wij zullen in deze paragraaf geval II beschouwen zonder eliminatie van $\underline{X}$ en $\underline{Y}$.

10.2. Notaties: Een waarnemingsresultaat $P_1, \ldots, P_n$ geven we aan met P , waarbij P het punt $(x_1, \ldots, x_n, y_1, \ldots, y_n)$ van een 2n-dimensionale ruimte $\mathcal{R}$ voorstelt. De bij $P_1, \ldots, P_n$ behorende punten $Q_1, \ldots, Q_n$ geven we tezamen aan met Q , waarbij Q het punt $(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ in een 2n-dimensionale ruimte $\mathcal{S}$ voorstelt.

$\underline{Q}$ bezit een waarschijnlijkheidsverdeling in $\mathcal{S}$, waarvan we de verdelingsfunctie

$$F(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$$

verkort aangeven door

$$F(Q)$$

---

1) $\overset{df}{=}$ betekent: is per definitie gelijk aan.

Tevens bezit $\underline{P}$, voor ieder punt $\mathcal{Q}$ van S, een voorwaardelijke waarschijnlijkheidsverdeling in R, onder de voorwaarde $\underline{\mathcal{Q}} = \mathcal{Q}$. De verdelingsfunctie van deze voorwaardelijke waarschijnlijkheidsverdeling:

$$G(x_1,\dots,x_n,y_1,\dots,y_n \mid \underline{X}_1 = X_1,\dots,\underline{X}_n = X_n, \underline{Y}_1 = Y_1,\dots,\underline{Y}_n = Y_n)$$

geven we verkort aan door $G(P|\mathcal{Q})$ en de onvoorwaardelijke verdelinsfunctie $G(x_1,\dots,x_n,y_1,\dots,y_n)$ van $\underline{P}$ in R door: $G(P)$. Wij hebben dan de volgende betrekking:

$$(50) \qquad G(P) = \int_S G(P|\mathcal{Q})\, dF(\mathcal{Q})$$

De in de vorige gedeelten afgeleide schattingen $\underline{a}$ (van α) en $\underline{b}$ (van β) zijn berekend onder de hypothese, dat $\mathcal{Q}$ een bepaald (zij het onbekend) punt van S is, en zij bezitten onder deze hypothese een waarschijnlijkheidsverdeling, waaruit hun eigenschappen zijn afgeleid. Hetzelfde geldt voor de stochastische grenzen van de voor α en β afgeleide betrouwbaarheidsintervallen. Wij zullen hier een schatting $\underline{a}$ van α, daar deze door P geheel bepaald is, aangeven door

$$\alpha(\underline{P})$$

en betrouwbaarheidsgrenzen voor α om dezelfde reden door

$$\underline{a}_1 = a_1(\underline{P}) \quad en \quad \underline{a}_2 = a_2(\underline{P})$$

10.3. Beschouwen wij nu de <u>methode der kleinste quadraten</u> voor het geval, dat de meetfout $\underline{u}$ in de x-richting identiek gelijk aan nul is, dan zien wij (zie 4.3), dat de voorwaarde van stelling 2, 3 en 4 voor ieder punt $\mathcal{Q}$ van S gelden, mits slechts de algemene eis, dat er onder $\mathcal{Q}_1,\dots,\mathcal{Q}_n$ minstens twee verschillende punten voorkomen, vervuld is. Wij onderstellen nu, dat de waarschijnlijkheidsverdeling van $\underline{\mathcal{Q}}$ in S zodanig is, dat de kans, dat hieraan voldaan is, gelijk is aan 1 en laten deze voorwaarde verder buiten beschouwing. Dan geldt:

<u>Stelling 6</u>: Stelling 2,3 en 4 blijven onveranderd geldig, indien men ξ_i vervangt door $\underline{X}_i$ $(i = 1,\dots,n)$, waarbij het punt $\underline{\mathcal{Q}} \equiv (\underline{X}_1,\dots,\underline{X}_n)$ een waarschijnlijkheidsverdeling in S bezit. ')

') $\underline{X}_1,\dots,\underline{X}_n$ behoeven dus niet onafhankelijk verdeeld te zijn.

Bewijs: Dit bestaat uit twee delen:

1) Indien voor een schatting $\underline{\alpha}$ van α in stelling 2, 3 of 4 bewezen is, dat zij zuiver is, betekent dit, dat nu geldt, dat

$$(51) \qquad \mathcal{E}_{\underline{\varphi}=\varphi}\,\underline{\alpha} = \int_R a(P)\,dG(P|\varphi) = \alpha$$

is voor iedere $\varphi \in S$.

Een dergelijke schatting is dus voorwaardelijk onder de voorwaarde $\underline{\varphi} = \varphi$, voor iedere $\varphi \in S$. echter ook onvoorwaardelijk zuiver, immers:

$$(52) \qquad \mathcal{E}\,\underline{\alpha} = \int_R a(P)\,dG(P) = \iint_{R\,S} a(P)\,dG(P|\varphi)\,dF(\varphi) =$$

$$= \int_S dF(\varphi) \int_R a(P)\,dG(P|\varphi) = \alpha \int_S dF(\varphi) = \alpha$$

Hetzelfde geldt voor de andere in stelling 2, 3 en noemde schattingen.

2) In stelling 4 is voor α een betrouwbaarheidsinterval afgeleid. Geven wij dit aan met $(\underline{\alpha}_1, \underline{\alpha}_2)$, dan geldt:

$$(53) \qquad P[\,\underline{\alpha}_1 < \alpha < \underline{\alpha}_2 \,|\, \underline{\varphi} = \varphi\,] = 1 - p$$

voor iedere $\varphi \in S$.

Derhalve is ook onvoorwaardelijk:

$$(54) \qquad P[\,a_1(P) < \alpha < a_2(P)\,] = \int_S dF(\varphi)\, P[\,\underline{\alpha}_1 < \alpha < \underline{\alpha}_2 \,|\, \underline{\varphi} = \varphi\,] =$$

$$= (1-p) \int_S dF(\varphi) = 1 - p.$$

Aanloog voor het betrouwbaarheidsinterval voor β en het betrouwbaarheidsgebied voor α en β gezamenlijk.

10.4. De overige in het vorige gedeelte vermelde resultaten laten zich niet zo volledig van geval I op geval II overbrengen, daar deze resultaten gebonden zijn aan voorwaarden betreffende φ, die in het algemeen niet voor iedere φ van S vervuld zijn. Deze voorwaarden, die wij nevenvoorwaarden zullen noemen, zijn:

Voorwaarde 4) van stelling 1 (zie 4.2);
voorwaarde 3 en 4 van Wald (zie 5.2);
voorwaarde 2 en 3 van Housner en Brennan (zie 6.2);
voorwaarde 1.c van 8.2;
voorwaarde 1 van 9.2.

Gedeeltelijk zijn deze voorwaarden gebruikt voor het afleiden van betrouwbaarheidsintervallen, gedeeltelijk voor het bewijs der bruikbaarheid van bepaalde schattingen. Voor ieder van deze methoden gelden nu de volgende beschouwingen:

1) Voor <u>betrouwbaarheidsintervallen</u>:

Indien V verzameling van alle φ uit S is, waarvoor aan de betreffende nevenvoorwaarde voldaan is, dan is een onder de hypothese $\underline{\varphi} = \varphi$ afgeleid betrouwbaarheidsinterval, met onbetrouwbaarheidsdrempel p, tevens een <u>voorwaardelijk</u> betrouwbaarheidsinterval, met dezelfde onbetrouwbaarheidsdrempel, onder de voorwaarde $\underline{\varphi} \in V$. (Bewijs analoog aan het tweede gedeelte van het bewijs van stelling 6). Is

(55) $$P[\underline{\varphi} \in V] = 1-q$$

dan is het betrouwbaarheidsinterval tevens een <u>onvoorwaardelijk</u> betrouwbaarheidsinterval met onbetrouwbaarheidsdrempel $\leqq p+q$ ').

2) Voor <u>bruikbaarheid</u>:

Aangezien dit een asymptotische eigenschap is, d.w.z. een limieteigenschap voor $n \to \infty$ schrijven wij S_n voor S en $\underline{\varphi}^{(n)}$ voor $\underline{\varphi}$. Als er nu een rij deelruimten V_n is, met $V_n \subset S_n$, zodanig, dat de betreffende nevenvoorwaarden <u>gelijkmatig</u> in n vervuld zijn voor $\varphi^{(n)} \in V_n$, geldt voor een onder de voorwaarden

$$\underline{\varphi}^{(n)} = \varphi^{(n)} \quad en \quad \varphi^{(n)} \in V_n$$

bruikbare schatting $\underline{a}_n$ van α:

(56) $$P[|a_n - \alpha| < \varepsilon \,|\, \underline{\varphi}^{(n)} = \varphi^{(n)}] > 1-\delta \quad voor \quad n > N(\varepsilon, \delta)$$

Is nu

(57) $$P[\underline{\varphi}^{(n)} \in V_n] = 1 - q_n$$

met

(58) $$\lim_{n\to\infty} q_n = 0$$

') Geheel analoog kan men bewijzen, dat de in 7.2 genoemde mediaan-zuiverheid van $\underline{a}_w$ en $\underline{a}_{HB}$ als schattingen van α voor geval II overgaat in een voorwaardelijke mediaanzuiverheid, onder de voorwaarde $\underline{\varphi} \in V$, terwijl onvoorwaardelijk geldt, dat $P[a<\alpha]$ voor beide schattingen tussen de waarden $\frac{1}{2}-q$ en $\frac{1}{2}+q$ ligt.

dan volgt (analoog aan de bewijsvoering van stelling 6)

(57) $P[|a_n - \alpha| < \varepsilon] > 1 - \delta - q_n$ voor $n > N(\varepsilon,\delta)$

zodat de schatting ook onvoorwaardelijk bruikbaar is. Vergelijk in dit verband ook WALD [17] punt 8.

III. Schatting van één coördinaat, als de andere gegeven is.

11. Het probleem .

11.1. Gegeven zijn: n punten $P_1, \ldots, P_n$ en de voorwaarden, die nodig zijn, om uit deze punten volgens één der beschreven methoden schattingen van α en β te bepalen. Verder is van een $(n+1)^{\text{e}}$ punt P_0 de coördinaat x_0 gegeven.

Gevraagd wordt een schatting te geven van de coördinaat η_0 (resp. y_0) van P_0 en van deze schatting de eigenschappen na te gaan. [1])

11.2. Geval II. In geval II bezitten de punten $P_1, \ldots, P_n$ en P_0 een simultane waarschijnlijkheidsverdeling, zodat men het probleem kan zien als de vraag naar de eigenschappen van de verdeling van y_0 en in het bijzonder van y_0, onder de voorwaarde $x_0 = x_0$, terwijl aan de coördinaten van $P_1, \ldots, P_n$ geen voorwaarden worden opgelegd. Dit is de beschouwingswijze der regressieanalyse.

11.3. Geval I. In geval I bezitten de punten $P_1, \ldots, P_n$ en P_0 eveneens een simultane waarschijnlijkheidsverdeling, nu echter met de coördinaten der punten $\xi_1, \ldots, \xi_n$ en ξ_0 als onbekende parameters. Men zou hier het probleem op dezelfde wijze kunnen stellen als onder geval II, voor zoverre het y_0 betreft; η_0 echter bezit geen voorwaardelijke waarschijnlijkheidsverdeling onder de voorwaarde $x_0 = x_0$ en in het algemeen is juist η_0 belangrijk (en niet y_0). Het ligt daarom voor de hand te trachten een schatting van η_0 te geven zonder de voorwaarde $x_0 = x_0$ te stellen.

Indien men het probleem voor de twee gevallen op deze twee verschillende wijzen interpreteert, treedt in het

---

[1]) De formulering met y_0 als gegeven coördinaat en te schatten abscis van P_0 ontstaat door verwisseling van de coördinaatassen uit de hier gegeven formulering.

geval van exact meetbare ξ resp. $\underline{X}$ een verschil in resultaat op, waarover in de litteratuur nogal wat te doen is geweest (zie b.v. A. WALD [17] p. 298 e.v., C. EISENHART [6] en D.V. LINDLEY [12] p. 231 e.v.). Wij zullen dit geval nader beschouwen:

12. Schatting van de abscis bij gegeven ordinaat, als de abscis foutloos is.

12.1. In het geval $\underline{u} \equiv 0$ kan men volgens de methode van 4.3 (stelling 2), indien de daar gemaakte onderstellingen vervuld zijn, met behulp van $\underline{P}_1, \ldots, \underline{P}_n$ zuivere schattingen $\underline{a}$ en $\underline{b}$ voor α resp. β afleiden, die dan (volgens stelling 6) voor beide gevallen zuiver zijn. Dan is echter (vgl. 7.1) in het algemeen $\frac{1}{\underline{a}}$ een onzuivere schatting van $\frac{1}{\alpha}$ en een zuivere schatting voor $\frac{1}{\alpha}$ is in dit geval niet bekend.

Indien nu van P_0 de coördinaat y_0 gegeven is, is dus

(59) $$\frac{1}{\underline{a}}(y_0 - \underline{b})$$

in geval I in het algemeen een onzuivere schatting van ξ_0, daar (in het algemeen)

(60) $$\xi_0 = \frac{1}{\alpha}(\eta_0 - \beta) \neq \mathcal{E}\,\frac{1}{\underline{a}}(y_0 - \underline{b})$$

is.

In geval II is (59) eveneens een onzuivere schatting van $\mathcal{E}_{y_0 = y_0} \underline{X}_0$, zelfs indien α en β bekend zijn en in (59) voor a en b worden gesubstitueerd. Dit blijkt op de volgende wijze:

Wij hebben (de index 0 voor het moment weglatende)

$$\underline{Y} = y - \underline{v}$$

waarin $\underline{Y}$ en $\underline{v}$ onafhankelijk zijn. Dus

$$\mathcal{E}_{y=y}\,\underline{Y} = \mathcal{E}_{y=y}\,y - \mathcal{E}_{y=y}\,\underline{v} = y - \mathcal{E}_{y=y}\,\underline{v}$$

Zij $h(y)$ verdelingsdichtheid van $\underline{Y}$ en $k(v)$ de verdelingsdichtheid van $\underline{v}$, dan is de simultane verdelingsdichtheid van $\underline{Y}$ en $\underline{v}$

$$h(y) \cdot k(v)$$

dus de simultane verdelingsdichtheid van y en $\underline{v}$ is:

$$h(y - v) \cdot k(v)$$

zodat wij hebben

$$(61) \qquad \mathcal{E}_{\underline{y}=y}\,\underline{v} = \frac{\int v\,h(y-v)\,k(v)\,dv}{\int h(y-v)\,k(v)\,dv}$$

Indien de teller de factor $h(y-v)$ niet bevatte, zou zij gelijk aan $\mathcal{E}\underline{v}$, dus gelijk aan nul zijn. Nu is dit in het algemeen niet het geval. Indien v klein is t.o.v. y , zodat in (61) in de Taylor-ontwikkeling van $h(y-v)$ de tweede- en hogere graadstermen verwaarloosd kunnen worden, krijgen we de volgende benadering

$$\mathcal{E}_{\underline{y}=y}\,\underline{v} \approx \frac{h(y)\int v\,k(v)\,dv - h'(y)\int v^2 k(v)\,dv}{h(y)\int k(v)\,dv - h'(y)\int v\,k(v)\,dv} =$$

$$= -\frac{h'(y)}{h(y)}\,\mathcal{E}\underline{v}^2$$

daar $\mathcal{E}\underline{v}=0$ is en $\int k(v)dv = 1$ is.

Indien b.v. $\underline{y}$ normaal verdeeld is met gemiddelde μ en spreiding σ en $\underline{v}$ normaal verdeeld met gemiddelde 0 en spreiding τ , krijgt men

$$(62) \qquad \mathcal{E}_{\underline{y}=y}\,\underline{v} = \frac{\tau^2}{\tau^2+\sigma^2}\,(y-\mu)$$

hetgeen slechts dan gelijk aan 0 is, als $\tau=0$ of $\mu=y$ is.

Nu is dus in het algemeen

$$(63) \qquad \mathcal{E}_{\underline{y}_0=y_0}\,\underline{X}_0 = \frac{1}{\alpha}\,\mathcal{E}_{\underline{y}_0=y_0}\,\underline{y}_0 - \frac{\beta}{\alpha} \neq \frac{1}{\alpha}\,(y_0-\beta)$$

Lindley geeft nu echter voor die gevallen, waarvoor stelling 5 geldt, een methode aan, om voor <u>geval</u> II een <u>zuivere</u> schatting van $\mathcal{E}_{\underline{y}_0=y_0}\,\underline{X}_0$ te verkrijgen. Indien n.l. stelling 5 geldt (wat b.v. het geval is, als $\underline{y}_0$ en $\underline{v}_0$ beide normaal verdeeld zijn, of indien beide een Γ -verdeling bezitten), is (daar $x_0 \equiv \underline{X}_0$ is wegens $\underline{\omega}_0 \equiv 0$):

$$(64) \qquad \mathcal{E}_{\underline{y}_0=y_0}\,\underline{X}_0 = \mathcal{E}_{\underline{y}_0=y_0}\,\underline{x}_0 = \frac{1}{\alpha'}\,y_0 - \frac{\beta'}{\alpha'}$$

Volgens de methode der kleinste quadraten te werk gaande (met minimalisering van de residuen in de richting van de <u>foutloze</u> coördinaat, dus i.c. de x -richting, verkrijgt men nu zuivere schattingen $\underline{c}'$ en $\underline{a}'$ van $\frac{1}{\alpha'}$ resp. van $-\frac{\beta'}{\alpha'}$, zodat

geldt:

(65) $\mathcal{E}_{\underline{y}_0=y_0}\,\underline{X}_0 = \frac{1}{\alpha'}y_0 - \frac{\beta'}{\alpha'} = \mathcal{E}_{\underline{y}_0=y_0}\,(\underline{\varsigma}'y_0 - \underline{\alpha}')$

De schatting

(66) $\underline{\varsigma}'y_0 - \underline{\alpha}'$

is nu dus een zuivere schatting van $\mathcal{E}_{\underline{y}_0=y_0}\,\underline{X}_0$.

Op te merken valt nog, dat deze methode in geval I geen oplossing geeft, daar in dat geval stelling 5 niet geldt, en $\underline{\varsigma}'$ een onzuivere schatting van $\frac{1}{\alpha}$ is. Dit laatste blijkt uit stelling 5; immers was $\underline{\varsigma}'$ een zuivere schatting van $\frac{1}{\alpha}$ in geval I, dan was dit volgens stelling 6 ook zo in geval II, hetgeen echter in strijd is met de tweede bij stelling 5 gemaakte opmerking. Voor geval I is een zuivere schatting voor $\frac{1}{\alpha}$ nog niet gevonden. Schatting (59) is, uit het oogpunt van bruikbaarheid, in geval I boven (66) te verkiezen, daar (59) in dit geval bruikbaar is en (66) in het algemeen niet (zie hiervoor 13).

12.2. Indien de exact meetbare coördinaat van P_0 gegeven is, zijn aan het zuiver schatten van de andere coördinaat in geen van beide gevallen moeilijkheden verbonden. Daar voor het geval, dat beide coördinaten meetfouten vertonen, geen zuivere schattingen van α en β bekend zijn, behoeft dit niet apart behandeld te worden. Wij wijzen er echter op, dat ook in deze situatie de voor het in 12.1 gestelde probleem door Lindley gegeven oplossing voor geval II van toepassing is, indien stelling 5 kan worden toegepast.

13. <u>Bruikbaarheid van schattingen van de ene coördinaat uit de andere</u>.

Het bruikbaarheidsbegrip wordt in dit verband gecompliceerder. Immers beschikt men over bruikbare schattingen $\underline{a}$ en $\underline{b}$ van α resp. β verkregen met behulp van de punten $P_1, \ldots, P_n$ dan betekent dit, dat $\underline{a}$ en $\underline{b}$ stochastisch naar α en β convergeren, voor $n \to \infty$. Om nu echter een bruikbare schatting te verkrijgen van b.v. de coördinaat η_0 (resp. Y_0) van P_0 moet men bovendien over een bruikbare schatting van de coördinaat ξ_0 (resp. X_0) van P_0 beschikken. Deze kan niet bestaan uit één enkele waarneming van $\underline{x}_0$, wel, eventueel, uit

het gemiddelde $\underline{\bar{x}}_0$ van m dergelijke waarnemingen. Is dit gemiddelde inderdaad een bruikbare schatting van ξ_0 (resp. χ_0), d.w.z. convergeert $\underline{\bar{x}}_0$ voor $m \to \infty$ stochastisch naar ξ_0 (resp. χ_0), dan is

(67) $$\underline{a}\,\underline{\bar{x}}_0 + \underline{b}$$

een <u>bruikbare</u> schatting van η_0 (resp. y_0) in die zin, dat deze uitdrukking stochastisch tot η_0 (resp. y_0) convergeert, als n en m <u>beide</u> naar ∞ gaan.

Daar nu voor bruikbaarheid geldt, dat, als $\alpha \neq 0$ en $\underline{a}$ een bruikbare schatting van α is, ook $\frac{1}{\underline{a}}$ een bruikbare schatting van $\frac{1}{\alpha}$ is, treden er hier verder weinig moeilijkheden op. Het is echter van belang er op te wijzen, dat in geval I met foutloze ξ, indien de in 12.1 gemaakte onderstellingen gelden, schatting (59) een <u>bruikbare</u> schatting van ξ_0 is, indien men daarin y_0 door een bruikbare schatting van η_0 vervangt en indien $\underline{a}$ en $\underline{b}$ bruikbare schattingen voor α en β zijn (vgl. stelling 1), terwijl dit met (66) in het algemeen niet het geval is, daar $\underline{c}'$ en $\underline{d}'$ in dit geval onzuivere schattingen van $\frac{1}{\alpha}$ en $-\frac{\beta}{\alpha}$ zijn.

14. <u>Keuze uit de verschillende methoden</u>.

14.1. Een algemeen principe voor deze keuze is, dat men meestal een beter resultaat verkrijgt, naarmate men meer onderstellingen gebruikt. Men kieze daarom die methode, waarbij men van zoveel mogelijk (gerechtvaardigde) onderstellingen gebruik maakt. Een volledig overzicht van de verschillende mogelijke onderstellingenstelsels met de daarbij behorende methoden en resultaten wordt, door zijn uitgebreidheid, onoverzichtelijk. Wij volstaan daarom met het aangeven van enkele belangrijke punten.

14.2. Is ξ foutloos ($\underline{u} \equiv 0$), dan is de methode der kleinste quadraten (zie 4.3; stelling 1, 2 en 3) onder zeer ruime verdere voorwaarden toepasbaar en soms zelfs de meest doeltreffende. Betrouwbaarheidsintervallen voor α en β verkrijgt men, voor zover bekend, met deze methode slechts, als $\underline{v}$ normaal verdeeld is.

14.3. Vertonen <u>beide coördinaten meetfouten</u>, dan verliest de methode der kleinste quadraten veel van zijn toepasbaarheid, daar nu geen betrouwbaarheidsintervallen voor α en β, vol-

gens deze methode afgeleid, bekend zijn, terwijl voor het verkrijgen van schattingen de verhouding der spreidingen $\sigma_{\underline{u}_i}$ en $\sigma_{\underline{v}_i}$ bekend moet zijn. Is echter deze verhouding bekend, terwijl verder van de verdelingen der punten $(\underline{u}_i, \underline{v}_i)$ slechts bekend is, dat $\underline{u}_i$ en $\underline{v}_i$ voor iedere i ongecorreleerd zijn en $\underline{u}_i$ en $\underline{v}_j$ voor iedere i en j eveneens, dan is genoemde methode de enige van alle besproken methoden, waarvan de voorwaarden vervuld zijn.

14.4. Bezitten de punten $(\underline{u}_i, \underline{v}_i)$ alle dezelfde verdeling, en zijn zij onafhankelijk van elkaar verdeeld, dan zijn de methoden van Wald, Housner en Brennan, Hemelrijk en beide methoden van Theil (onder een aantal vrij algemene aanvullende voorwaarden) alle toepasbaar. Hiermee zijn schattingen en betrouwbaarheidsintervallen voor α, β, $\sigma_{\underline{u}}$ en $\sigma_{\underline{v}}$ te verkrijgen. De methode van Wald is de enige, waarmee onder deze omstandigheden schattingen van $\sigma_{\underline{u}}$ en $\sigma_{\underline{v}}$ verkregen worden. De methoden van Hemelrijk en Theil leveren betrouwbaarheidsintervallen, ook als de meetfouten niet normaal verdeeld zijn. Zijn deze wel normaal verdeeld, dan verkrijgt men ook met de methode van Wald betrouwbaarheidsintervallen.

14.5. Voor keuze tussen de methoden van Wald, Housner en Brennan, Hemelrijk en de beide methoden van Theil is (in verband met de nevenvoorwaarden voor $\mathcal{Q}$, de volgorde der punten $\mathcal{Q}_i$ betreffende) de vorm van de puntenwolk van invloed. Is deze "halter-vormig" (twee punt-wolken met een tussenruimte) dan is de methode van Wald aan te bevelen; heeft zij de vorm van een "bol met twee uitsteeksels", dan die van Hemelrijk; heeft zij de vorm van een staaf, dan die van Housner en Brennan (waarmee slechts schattingen verkregen worden) en de twee methoden van Theil. Indien slechts weinig punten beschikbaar zijn kunnen de methoden van Hemelrijk en de tweede van Theil toch gebruikt worden voor de bepaling van een betrouwbaarheidsinterval voor α met niet te grote onbetrouwbaarheidsdrempel p (resp. 7 en 5 voor $p = 0{,}05$).

14.6. Tenslotte zij er nogmaals op gewezen, dat het verrichten van (liefst alle) waarnemingen in duplo (of nog vaker) van groot nut is voor het verkrijgen van een scherpere oplossing, daar men dan 1^e met de gemiddelden van deze stelsels kan werken (als van ieder punt $\mathcal{Q}_i$ evenveel waarnemingen

zijn verricht), terwijl nu bovendien de normaliteit der fouten kan toetsen en hun spreidingen kan schatten.

15. Generalisaties van de theorie.

Voor meerdere variabelen ($\eta = \sum \alpha_i \xi_i + \beta$) vindt men een behandeling met de methode der kleinste quadraten en de maximum-likelihood methode bij LINDLEY [12] . Zie ook H. THEIL [16] , waarin tevens een litteratuurlijst zal worden opgenomen van publicaties betreffende stelsels van lineaire vergelijkingen van stochastische variabelen. H. THEIL [16] en K.R. NAIR and M.R. SHRIVASTAVA [14] behandelen ook het geval, dat η gelijk is aan een polynoom van ξ . De toepassing van de methode der kleinste quadraten op dit probleem wordt o.a. behandeld door W.E. DEMING [4].

16. Lijst van geciteerde litteratuur.

[1] M.S. Bartlett, Fitting a straight line when both variables are subject to error, Biometrics 5 (1949) 207-212.

[2] H. Cramér, Mathematical methodes of Statistics, Princeton 1946, p. 548 e.v.

[3] D. van Dantzig, Kadercursus Mathematische Statistiek, Math. Centrum 1947-1949. Hoofdstuk 4, p. 194 e.v.

[4] W.E. Deming, Some notes on least squares, Washington, Dept. of Agriculture 1938.

[5] F.N. David and J. Neyman, Extension of the Markoff theorem on least squares, Statistical research memoirs 2 (1938) 105-116.

[6] C. Eisenhart, The interpretation of certain regression methods and their use in biological and industrial research, Am. Math. Stat. 10 (1939) 162-186.

[7] J. Hemelrijk, Construction of a confidence region for a line, Proc. Kon. Ned. Ak. 52 (1949) 995-1005.

[8] G.W. Housner and J.F. Brennan, The estimation of linear trends, Ann Math. Stat. 19 (1948) 380-393.

[9] R.S. Koshal, J. Roy. Stat. Soc. 96 (1933) 303.

[10] M.G. Kendall, The advanced theory of Statistics, I, London 1947, Ch. 16, p. 388 e.v.

[11] M.G.Kendall, Rank correlation methods, London 1948.

[12] D.V. Lindley, Regression lines and the linear functional relationship, Suppl.Jrn.Roy.Stat.Soc. 9 (1947) 218-244.

[13] A.A. Markoff, Wahrscheinlichkeitsrechnung, Leipzig & Berlin 1912.

[14] K.R. Nair and M.P. Shrivastava, On a simple method of curve fitting, Sankhya 6 (1942) 121.

[15] K.R. Nair and K.S. Banerjee, A note on fitting of straight lines if both variables are subject to error, Sankhya 6 (1942) 331.

[16] H. Theil, A method of linear and polynomial ordered regression analysis. (Verschiint binnenkort in de Proc. Kon. Ned.Ak.).

[17] A. Wald, The fitting of straight lines if both variables are subject to error, Am.Math.Stat. 11 (1940) 284-300.