STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

SP 92
(S 344)

W. Molenaar

Overzicht van ontmengingsmethoden voor twee normale verdelingen.

(Statistica Neerlandica, 19(1965) p. 249-263)

Overzicht van ontmengingsmethoden voor twee normale verdelingen *)

door W. Molenaar **)

UDC 519.2 + 518.5

Summary

The parameters of a mixture of two normal distributions can be estimated in several ways from a sample of grouped observations. Various methods, already known in the literature, are listed and compared.

After an introduction and a statement of the problem, moment estimators, graphical methods and maximum likelihood estimators are mentioned. Some remarks are made on the advantages of the various methods, and on classification. The results of the methods are compared in five representative examples of mixtures.

Inleiding

Het is bij veel statistisch werk gebruikelijk, een gegeven waarnemingsreeks te interpreteren als een serie aselecte en onafhankelijke trekkingen uit een onbekende kansverdeling. Als men de aard (analytische vorm) van deze verdeling heeft vastgelegd, zal men proberen de parameters uit de waarnemingsreeks te schatten. Zo neemt men dikwijls – met of zonder voldoende motivering – aan, dat de verdeling normaal is, en men schat vervolgens verwachting en variantie.

In diverse situaties is het redelijk, aan te nemen dat een aselecte trekking heeft plaats gevonden uit een mengsel van twee verdelingen van dezelfde (bekende) aard. In dit geval kan men uit de steekproef de mengverhouding en de parameters van de beide componenten schatten. De hierbij gebruikte technieken kan men ruwweg indelen in grafische methoden, momentenmethoden en meest aannemelijke schatters. In de literatuur vindt men deze drie soorten van ontmenging o.a. beschreven voor normale, exponentiële, binomiale, Poisson- en Weibull-verdelingen. De grafische methoden zijn ook bruikbaar voor meer dan twee componenten, mits er niet te veel overlapping optreedt; men behoeft het aantal componenten niet tevoren vast te leggen. Uitbreiding van de andere methoden, tot een vast aantal componenten, is in principe mogelijk. Maar bij deze methoden wordt het oplossen van de onbekenden een

*) Lezing gehouden op de Statistische Dag 1965; op enige kleine aanvullingen na is deze tekst dezelfde als die van Rapport S 344 van de Statistische Afdeling van het Mathematisch Centrum.

**) Mathematisch Centrum, Amsterdam.

hachelijke zaak, door het vele rekenwerk, maar vooral door de afwijkingen tussen waargenomen en verwachte aantallen per klasse.

In dit rapport wordt uitsluitend gesproken over het ontmengen van twee normale verdelingen. Verder is verondersteld dat de gegeven waarnemingen *gegroepeerd* zijn in klassen van breedte b. Het is geen beperking $b = 1$ te stellen.

Notaties, formulering van het probleem

Gegeven: voor $k = 0, 1, \ldots, K$ vallen f_k waarnemingen in het interval $(a + k - \frac{1}{2}, a + k + \frac{1}{2})$; deze waarnemingen zijn een aselecte steekproef uit een mengsel van twee normale verdelingen met verdelingsfunctie G en verdelingsdichtheid

$$g(x) = \frac{\tau}{\sigma_1\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\} + \frac{1-\tau}{\sigma_2\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right\}. \qquad (1)$$

Gevraagd: schat de vijf parameters τ, μ_1, μ_2, σ_1 en σ_2. Het gestelde probleem is eenduidig oplosbaar: bij een dichtheid van de vorm (1) behoort niet meer dan één parametervector, zoals blijkt uit een stelling van TEICHER [16].

Wij voeren nog in

$$n \overset{\text{def}}{=} \sum_{k=0}^{K} f_k \quad \text{en} \quad x_k \overset{\text{def}}{=} a + k \ (k = 0, 1, \cdots, K). \qquad (2)$$

Momentenmethoden

KARL PEARSON [11] stelt de eerste vijf momenten van de verdeling met dichtheid (1) gelijk aan die van de steekproef. Eliminatie in het stelsel van vijf vergelijkingen met vijf onbekenden leidt tot een vergelijking van de negende graad.

BURRAU en STRÖMGREN [3, 15] gebruiken de eerste vijf kumulanten; via diagrammen en tabellen vindt men twee hulpgrootheden waarmee men schattingen van de vijf parameters kan berekenen.

RAO [13] gebruikt in het geval van gelijke varianties een methode van CHARLIER [4], die gebruik maakt van de eerste vier steekproefkumulanten $k_i (i = 1, \ldots, 4)$. Men moet dan, b.v. door iteratie, de (unieke) oplossing met $-k_2 < x < 0$ zoeken van $2x^3 + k_4 x + k_3^2 = 0$. Als nu d_1 de negatieve en d_2 de positieve wortel is van $d^2 + k_3 d/x + x = 0$, dan is $t = d_2/(d_2 - d_1)$, $m_1 = k_1 + d_1$, $m_2 = k_1 + d_2$ en $s^2 = k_2 + x$ een schatting voor de parameters.

In het geval van gelijke varianties kan men ook een hulptabel van STRÖMGREN [15] gebruiken, of de diagrammen van PRESTON [12] of SITTIG [14]. De laatste geeft ook diagrammen voor het geval van gelijke mengverhouding (met eventueel ongelijke varianties). Een momentenmethode voor het geval van gelijke (eventueel onbekende) verwachtingen vindt men bij AGARD [1]. Als beide

verwachtingen bekend en ongelijk zijn is er een methode van CHARLIER [4], ook beschreven door COURT [5].

Grafische methoden

VAN ALPHEN [2] beveelt aan de frequentiecurve voor het mengsel met het „timmermansoog" te splitsen, en daarbij als hulpmiddel scharen van normale curven te vervaardigen.

HALD [10] merkt op dat voor één normale verdeling, dus met $\tau = 1$ in (1), geldt

$$f_k \approx \frac{n}{\sigma_1\sqrt{2\pi}} \exp\left\{-\frac{(x_k - \mu_1)^2}{2\sigma_1^2}\right\}, \tag{3}$$

mits de klassebreedte niet te groot is t.o.v. de standaardafwijking en de waargenomen frequentie f_k niet te veel afwijkt van de verwachte frequentie. Voor één normale verdeling is de grafiek van $\log f_k$ als functie van x_k dus nagenoeg een parabool met as $x = \mu_1$ en vorm bepaald door σ_1. Zet men nu voor het mengsel f_k uit tegen x_k, op halflogarithmisch papier, dan kan men deze grafiek in twee parabolen splitsen; meestal is aan één der uiteinden de ene component veel sterker vertegenwoordigd dan de andere. Uit de parabolen bepaalt men μ_i en σ_i^2; zij leveren tevens voor elke k een splitsing $f_k = f_k(1) + f_k(2)$, d.w.z. $f_k(i)$ waarnemingen x_k in de i^e component ($i = 1, 2$). Men schat de mengverhouding nu met $t = \sum_k f_k(1)/n$.

DAEVES en BECKEL [6] merken op, dat het uitzetten van f_k/n tegen x_k op (cumulatief) waarschijnlijkheidspapier voor één normale verdeling leidt tot een hyperbool-achtige curve, die symmetrisch is t.o.v. de verwachting. De grafiek van het mengsel kan op grond van die symmetrie in twee dergelijke curven worden gesplitst. Men beschikt dan weer over $f_k(1)$ en $f_k(2)$, en de mengverhouding wordt geschat zoals bij HALD. Voor de afzonderlijke waarnemingsreeksen $\{x_k, f_k(1)\}$ en $\{x_k, f_k(2)\}$ kan men de verwachtingen en standaardafwijkingen grafisch schatten, door de cumulatieve verdelingsfuncties uit te zetten op waarschijnlijkheidspapier: de daarbij te vinden rechte lijnen snijden het 50% niveau in $x = \mu_i$ en het 15,9% niveau in $x = \mu_i - \sigma_i$ ($i = 1, 2$).

WEICHSELBERGER [17] vindt voor één normale verdeling, dus $\tau = 1$ in (1), uit (3) de relatie

$$y \overset{\text{def}}{=} \log\frac{f_{k+1}}{f_k} \approx \frac{(a + k - \mu_1)^2 - (a + k + 1 - \mu_1)^2}{2\sigma_1^2} = Ak + B = Ax + C, \tag{4}$$

waar A, B en C geschikte constanten zijn. De grafiek van f_{k+1}/f_k tegen x_k op halflogarithmisch papier is voor $\tau = 1$ dus bij benadering een rechte lijn, die het niveau $y = \log 1$ snijdt in $x = \mu_1$ en het niveau $y = \log e$ in $x = \mu_1 - \sigma_1$.

Voor een mengsel kan men in de grafiek meestal de twee samenstellende lijnen onderscheiden en daaruit μ_i en σ_i schatten. Als men in de grafiek een hulplijn trekt, kan men na enig rekenwerk ook de mengverhouding schatten.

ESSENWANGER [7, 8, 9] past met de methode der kleinste kwadraten een normale dichtheid aan bij een aantal punten (x_k, f_k) en splitst de zo verkregen component af.

Meest aannemelijke schatters

De volgende methode om iteratief de meest aannemelijke schatters te vinden is voor gelijke varianties beschreven door RAO [13]. Als $\theta = (\tau, \mu_1, \mu_2, \sigma_1, \sigma_2)$ de parametervector is en $\underline{x}$ de stochastische variabele met verdelingsdichtheid (1), dan voert men in:

$$P_k(\theta) = P\left\{\underline{x} \in \left(x_k - \frac{1}{2}, x_k + \frac{1}{2}\right) \mid \theta\right\} \qquad (k = 0, 1, \cdots, K);$$

$$\log L(\theta) = \sum_{k=0}^{K} f_k \log P_k(\theta); \tag{5}$$

$$S_i(\theta) = \frac{\partial \log L(\theta)}{\partial \theta_i} = \sum_k \frac{f_k}{P_k(\theta)} \frac{\partial P_k(\theta)}{\partial \theta_i} \quad (i = 1, 2, \cdots, 5).$$

Voor de meest aannemelijke schatting $\hat{\theta}$ geldt $S_i(\hat{\theta}) = 0$ voor $i = 1, 2, \ldots, 5$. Laat θ° een grafische of momentenschatting zijn en stel $\mathrm{d}\theta = \hat{\theta} - \theta^\circ$. Bij verwaarlozing van $|\, \mathrm{d}\theta \,|^2$ t.o.v. $|\, \mathrm{d}\theta \,|$ en van

$$\frac{\partial P_k(\hat{\theta})}{\partial \theta_i} - \frac{\partial P_k(\theta^0)}{\partial \theta_i} \text{ t.o.v. } \frac{\partial P_k(\hat{\theta})}{\partial \theta_i} \text{ vindt men:}$$

$$P_k(\hat{\theta}) \approx P_k(\theta^0) + \sum_{j=1}^{5} \mathrm{d}\theta_j \frac{\partial P_k(\theta^0)}{\partial \theta_j};$$

$$\frac{1}{P_k(\hat{\theta})} \approx \frac{1}{P_k(\theta^0)} \left\{1 - \sum_j \mathrm{d}\theta_j \frac{1}{P_k(\theta^0)} \frac{\partial P_k(\theta^0)}{\partial \theta_j}\right\};$$

$$0 = S_i(\hat{\theta}) = \sum_k \frac{f_k}{P_k(\hat{\theta})} \frac{\partial P_k(\hat{\theta})}{\partial \theta_i} \approx \tag{6}$$

$$\approx \sum_k \frac{f_k}{P_k(\theta^0)} \frac{\partial P_k(\theta^0)}{\partial \theta_i} - \sum_k \sum_j \mathrm{d}\theta_j \frac{f_k}{P_k^2(\theta^0)} \frac{\partial P_k(\theta^0)}{\partial \theta_i} \frac{\partial P_k(\theta^0)}{\partial \theta_j} =$$

$$= S_i(\theta^0) - \sum_{j=1}^{5} I_{ij}(\theta^0)\, \mathrm{d}\theta_j,$$

waar

$$I_{ij}(\theta) \overset{\text{def}}{=} \sum_{k=0}^{K} \frac{f_k}{P_k^2(\theta)} \frac{\partial P_k(\theta)}{\partial \theta_i} \frac{\partial P_k(\theta)}{\partial \theta_j}.$$

Nu geldt, als

$$a_{i,k} = \frac{x_k + \frac{1}{2} - \mu_i}{\sigma_i} \quad \text{en} \quad \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\tfrac{1}{2} u^2\right) \mathrm{d}u,$$

$$P_k(\theta) = \int_{x_k - \frac{1}{2}}^{x_k + \frac{1}{2}} g(x)\,\mathrm{d}x = \tag{7}$$

$$= \tau\Phi(a_{1,k}) - \tau\Phi(a_{1,k-1}) + (1-\tau)\,\Phi(a_{2,k}) - (1-\tau)\,\Phi(a_{2,k-1}),$$

waaruit $\dfrac{\partial P_k}{\partial \theta_i}$ voor $i = 1, 2, \cdots, 5$ direct volgen.

Voor de Electrologica $X1$ rekenmachine van het Mathematisch Centrum is een ALGOL-programma geschreven met als invoer de paren $x_k, f_k (k = 0, 1, \ldots, K)$, een beginschatting $\theta^\circ = (t^\circ, m_1{}^\circ, m_2{}^\circ, (s_1{}^\circ)^2, (s_2{}^\circ)^2)$ en vier constanten e_1, e_2, e_3 en M. Voor $m = 0, 1, \ldots$ worden achtereenvolgens berekend:

$$\frac{\partial P_k(\theta^m)}{\partial \theta_i}, \ \log L(\theta^m), \ S_i(\theta^m) \text{ en } I_{ij}(\theta^m) \quad (i, j = 1, \cdots, 5);$$

$$\mathrm{d}\theta_j^m \text{ uit } \sum_{j=1}^{5} I_{ij}(\theta^m)\,\mathrm{d}\theta_j^m = S_i(\theta^m) \qquad (i = 1, \cdots, 5); \tag{8}$$

$$\theta^{m+1} = \theta^m + \mathrm{d}\theta^m.$$

De iteratie wordt gestaakt zodra òf M stappen gedaan zijn, òf één van de voorwaarden

$$\left| \frac{\log L(\theta^m) - \log L(\theta^{m-1})}{\log L(\theta^m)} \right| < 10^{-e_2} \tag{9}$$

of

$$\left| \frac{\mathrm{d}\theta_i^m}{\theta_i^m} \right| < 10^{-e_3} \text{ voor elke } i \tag{10}$$

vervuld is. Zodra $| 1 - s_1/s_2 | < 10^{-e_1}$ is, wordt $s_1 = s_2$ gesteld en verder gewerkt met vier vergelijkingen met vier onbekenden. Als de iteratie, zeg na m stappen, gestaakt is, produceert de machine de schattingen voor de parameters, de geschatte standaardafwijkingen van deze schattingen en de bij de geschatte parametervector behorende kansen. De uitvoer bestaat nl. uit

$$t^h, m_1^h, m_2^h, s_1^h, s_2^h, \log L(\theta^h) \ (h = 1, 2, \cdots, m);$$

$$\sqrt{I_{ii}^{-1}(\theta^m)} \approx \sigma(\theta_i^m) \quad (i = 1, 2, \cdots, 5);$$

$$P\{\underline{x} \leqq a - \tfrac{1}{2} \mid \theta^m\}; \tag{11}$$

$$P_k(\theta^m) \quad \text{voor} \quad k = 0, 1, \cdots, K;$$

$$P\{\underline{x} > a + K + \tfrac{1}{2} \mid \theta^m\}.$$

Vergelijking der ontmengingsmethoden

Bij de keuze van een methode om het besproken ontmengingsprobleem op te lossen zal men zich o.a. laten leiden door de aard en omvang van de waarnemingsreeks, de beschikbare tijd en rekenapparatuur en de vereiste nauwkeurigheid. De volgende opmerkingen over bezwaren en voordelen van de genoemde methoden kunnen daarom slechts aarzelende aanwijzingen vormen voor de practicus, die bij ieder ontmengingsprobleem zijn eigen wensen en mogelijkheden als voornaamste leidraad zal moeten beschouwen.

De *grafische methoden* van HALD en DAEVES & BECKEL zijn beiden weinig tijdrovend en reeds voor steekproeven van enige honderden waarnemingen redelijk nauwkeurig. Wie eenmaal een verzameling papieren mallen $x^2 = 2py$ voor diverse p-waarden vervaardigd heeft, zal vermoedelijk aan HALD de voorkeur geven: daar is één grafiek voldoende, terwijl de andere methode ook nog het cumulatief uitzetten van de afzonderlijke componenten vereist. Het op het oog aanpassen van twee normale curven (VAN ALPHEN) eist het vervaardigen van een schaar curven met twee parameters en een begaafd timmermansoog. De verschillen tussen waargenomen en verwachte frequenties worden bij het vormen van quotiënten zo geaccentueerd, dat WEICHSELBERGER's methode alleen bij vele duizenden waarnemingen een redelijk resultaat oplevert. In een voorbeeld van deze auteur telt elke component 6000 waarnemingen; in onze voorbeelden van 400, 1000 of 2000 waarnemingen voor beide componenten samen, leidt de methode tot tamelijk dwaze uitkomsten. De methode van ESSENWANGER eist formidabel rekenwerk, zelfs bij gebruik van de in [9] gegeven tabellen.

Alle *momentenmethoden* hebben het bezwaar, dat het vierde en vooral het vijfde moment uit kleine steekproeven niet nauwkeurig kan worden geschat. De weinige waarnemingen die ver van het steekproefgemiddelde liggen geven hier de doorslag. In onze voorbeelden 1, 3 en 5 wijkt de oplossing volgens BURRAU dan ook sterk af van de overige schattingen; in de voorbeelden 2 en 4 leidt de methode tot een negatieve uitkomst voor de variantie. De combinatie van veel rekenwerk en het tussentijds gebruik van diagrammen is ook niet erg aantrekkelijk. Zodra men het aantal onbekenden tot vier kan verminderen door gelijke varianties of gelijke mengverhoudingen te veronderstellen, worden momentenmethoden veel attractiever. De oplossing van RAO's derdegraadsvergelijking (bij gelijke varianties) kan men vermijden door BURRAU's tabel, PRESTON's diagram of SITTIG's diagrammen te gebruiken; het werken met het diagram van PRESTON is het eenvoudigste. In de voorbeelden verschillen de resultaten van de vier methoden meestal maar weinig; men ziet daar tevens dat de uitkomsten nog tamelijk bevredigend kunnen zijn, ook als de varianties niet onaanzienlijk verschillen. SITTIG's methode voor mengverhouding $\frac{1}{2}$ blijkt

TABEL 1

Steekproefaantallen, parametervectoren, waarnemingen.

	Vb 1	Vb 2	Vb 3	Vb 4	Vb 5
n	1000	400	2000	1000	1000
τ	0,386	0,41	0,696	0,35	0,5
μ_1	—1,03	—1,75	—1	—2,2	0
μ_2	0,95	2,3	0,5	3,6	0
σ_1	5	1,25	1,75	1,2	1
σ_2	3,3	1,3	1,87	3,8	4
$\bar{\sigma}$ [1])	3,895	1,274	1,807	1,608	1,372
$(\mu_2 - \mu_1)/\bar{\sigma}$	0,51	3,18	0,83	3,61	0
σ_1/σ_2	1,52	0,96	0,94	0,32	0,25
x_k	f_k	f_k	f_k	f_k	f_k
—13	2				1
—12	0				0
—11	2				1
—10	4				5
—9	13				2
—8	14				3
—7	19		1	2	11
—6	21		5	6	10
—5	44	1	27	15	12
—4	60	17	78	45	28
—3	70	36	198	114	36
—2	72	36	337	147	60
—1	86	56	430	113	185
0	91	35	344	61	259
1	102	58	283	61	197
2	107	64	161	67	68
3	77	57	94	73	32
4	68	33	30	68	22
5	48	7	10	52	24
6	52		1	45	18
7	16		1	42	11
8	13			29	10
9	8			30	3
10	6			9	0
11	3			7	1
12	2			11	1
13				3	
	1000	400	2000	1000	1000

[1]) $\bar{\sigma}$ is gedefiniëerd door de relatie $\bar{\sigma}^{-1} = \sqrt{(\sigma_1^{-2} + \sigma_2^{-2})/2}$.

nog matig goede uitkomsten te geven, ook als de mengverhouding 0,4 of zelfs 0,7 is; in voorbeeld 5 waar werkelijk $\tau = \frac{1}{2}$ is levert de methode niets op omdat het vierde moment te groot is. De methode van AGARD voor gelijke verwachtingen werkt bevredigend, maar is nogal gevoelig voor afwijkingen tussen steekproefgemiddelde en verwachting.

De iteratieve bepaling van de *meest aannemelijke schatters* verloopt alleen naar wens, als de beginschatting al vrij dicht bij de meest aannemelijke waarde ligt. Vanuit een ander beginpunt vindt men misschien een relatief maximum van log L, en nog waarschijnlijker is dat het iteratieproces gaat divergeren; dit laatste gebeurde o.a. in voorbeeld 1 vanuit de exacte gegevens (!) en in voorbeeld 2 vanuit HALD's schatting. De genoemde verwaarlozingen waren in die gevallen blijkbaar te grof. In de overige gevallen waren na vier interatiestappen log L tot in 8 à 10 en de parameterwaarden tot in 2 à 3 significante cijfers constant. Het proces vergt vrij veel rekentijd, o.a. omdat de normale verdelingsfunctie voor veel argumenten moet worden berekend. Een grovere groepering van de waarnemingen vermindert de rekentijd, maar betekent natuurlijk verlies van informatie.

Classificatie

Als men na afloop van een ontmenging nog een nieuwe waarneming $\underset{\sim}{x} = x$ trekt, kan men zich afvragen aan welke component deze moet worden toegeschreven. Het antwoord wordt bepaald door de *locale mengverhouding* : het aandeel van de tweede component in het mengsel is in het punt x

$$\varrho(x) = \frac{\dfrac{1-\tau}{\sigma_2\sqrt{2\pi}} \exp\left\{-\dfrac{(x-\mu_2)^2}{2\sigma_2^2}\right\}}{g(x)}. \tag{12}$$

Men kan direct afleiden

$$\operatorname{sgn} \varrho'(x) = \operatorname{sgn}\left\{\frac{x-\mu_1}{\sigma_1^2} - \frac{x-\mu_2}{\sigma_2^2}\right\}. \tag{13}$$

Voor gelijke varianties en $\mu_2 > \mu_1$ zal de functie ϱ monotoon stijgen van 0 tot 1. Voor ongelijke varianties overheerst aan beide buitenzijden de component met de grootste variantie. De figuren 1 en 2 geven $\varrho(x)$ voor de parametervectoren „ML v/a H. of R." en „ML v/a B.I" van voorbeeld 1 (zie tabel 2), die ongeveer even aannemelijk zijn en resp. gelijke en ongelijke varianties hebben. De figuren zijn door de rekenmachine vervaardigd, met een zgn. plotter.

Vooral bij gelijke varianties en grote afstand tussen de verwachtingen kan het zin hebben, alle waarnemingen uit het gebied G_1 waar $\varrho(x) < 0{,}05$ is aan de eerste component toe te schrijven en alle waarnemingen uit G_2 ($\varrho(x) > 0{,}95$)

TABEL 2

Resultaten van de ontmengingen.

	τ	μ_1	σ_1	μ_2	σ_2	log L	χ^2	vg	P_{χ^2}
VOORBEELD 1									
exact	0,386	—1,03	5	0,95	3,3	—2822,521	27,01	21	0,17
ML v/a B.I	0,878	—0,03	4,14	2,67	2,23	—2817,889	20,17	15	0,17
st.afw.	0,110	0,31	0,14	0,95	0,99				
ML v/a H. of R.	0,245	—3,28	3,50	1,46	3,50	—2818,079	20,96	16	0,18
st.afw.	0,126	0,85	0,25	0,60	0,25				
Rao =	0,244	—3,213	3,526	1,432	3,526	—2818,089	20,97	16	0,18
Burrau =	0,244	—3,213	3,526	1,431	3,526	—2818,089	20,97	16	0,18
Preston =	0,22	—3,30	3,564	1,33	3,564	—2818,120	21,06	16	0,18
Hald	0,280	—2,7	3,61	1,7	3,61	—2819,443	22,84	16	0,12
1 Normale		0,3	4,050			—2820,255	25,36	19	0,14
Burrau I	0,785	—0,124	4,101	2,907	2,443	—2820,275	24,90	15	0,051
Burrau II	0,789	—0,122	4,094	2,995	2,396	—2820,520	25,32	15	0,044
D & B	0,307	—3,7	2,8	1,9	3,6	—2828,275	38,11	16	0,001
Sittig $\frac{1}{2}$	0,5	—2,048	3,870	2,648	2,609	—2830,561	45,59	16	0,0001
Sittig =	0,21	—5,071	2,956	1,728	2,956	—2832,941	51,07	16	2.10^{-5}
VOORBEELD 2									
exact	0,41	—1,75	1,25	2,30	1,30	—888,043	12,27	9	0,20
ML v/a R.	0,416	—1,772	1,305	2,181	1,305	—886,933	9,38	4	0,052
st.afw.	0,031	0,137	0,077	0,114	0,077				
Sittig =	0,40	—1,863	1,290	2,138	1,290	—887,234	10,28	5	0,067
Rao =	0,415	—1,834	1,231	2,220	1,231	—887,814	12,26	5	0,032
Preston =	0,417	—1,828	1,229	2,227	1,229	—887,866	12,39	5	0,030
Burrau =	0,414	—1,839	1,227	2,221	1,227	—887,927	12,55	5	0,028
Sittig $\frac{1}{2}$	0,5	—1,437	1,497	2,512	0,987	—891,101	17,73	4	0,001
Hald	0,30	—2,3	1,04	2,0	1,81	—897,183	26,35	5	0,0001
1 Normale		0,5375	2,346			—911,717	54,71	9	$<10^{-5}$
Weichs. II	0,61	—1,7	1,55	2,1	1,7	—918,997	74,04	5	$<10^{-5}$
Weichs. I	0,48	—2,7	0,84	2,1	1,7	—941,764	124,11	4	$<10^{-5}$
VOORBEELD 3									
exact	0,696	—1,00	1,75	0,50	1,87	—4179,227	12,56	11	0,32
ML v/a H. of R.	0,740	—1,204	1,601	1,285	1,601	—4175,968	5,90	7	0,55
st.afw.	0,055	0,139	0,065	0,212	0,065				
Sittig =	0,75	—1,193	1,593	1,356	1,593	—4176,038	6,05	7	0,53
Burrau =	0,761	—1,139	1,637	1,295	1,637	—4176,101	6,23	7	0,51
Rao =	0,761	—1,138	1,637	1,295	1,637	—4176,102	6,23	7	0,51
Preston =	0,765	—1,119	1,651	1,275	1,651	—4176,222	6,49	7	0,48
Sittig $\frac{1}{2}$	0,5	—1,487	1,552	0,375	1,836	—4176,301	6,72	7	0,46
Hald	0,705	—1,35	1,50	1,30	1,47	—4177,647	8,85	6	0,18
Burrau	0,286	—1,732	1,232	—0,262	1,971	—4181,123	16,65	6	0,010
D & B	0,765	—1,19	1,63	1,15	1,47	—4182,306	7,15	6	0,31
1 Normale		—0,556	1,938			—4183,573	20,82	9	0,014
Weichs.	0,722	—1,2	1,64	0,4	1,70	—4198,487	57,87	5	$<10^{-5}$

TABEL 2 vervolg

	τ	μ_1	σ_1	μ_2	σ_2	log L	χ^2	vg	P_{χ^2}
VOORBEELD 4									
exact	0,35	—2,20	1,20	3,60	3,80	—2710,731	21,60	19	0,30
ML v/a H. of D.	0,354	—2,197	1,117	3,214	3,818	—2707,978	11,43	13	0,58
st.afw.	0,031	0,089	0,093	0,266	0,155				
D & B	0,342	—2,25	1,05	3,15	3,95	—2708,828	17,56	14	0,23
Hald	0,361	—2,2	1,14	3,2	3,61	—2709,690	19,48	13	0,11
Sittig $\frac{1}{2}$	0,5	—1,818	1,639	4,410	3,319	—2719,495	38,96	14	0,0004
Preston =	0,73	—0,768	2,246	6,869	2,246	—2772,628	155,95	14	$<10^{-5}$
Sittig =	0,73	—0,766	2,247	6,872	2,247	—2772,664	155,95	14	$<10^{-5}$
Weichs.	0,447	—2,05	1,28	2,28	2,95	—2804,235	339,70	9	$<10^{-5}$
1 Normale		1,296	4,078			—2824,620	257,78	18	$<10^{-5}$
VOORBEELD 5									
exact	0,5	0	1	0	4	—2275,991	28,35	16	0,029
ML v/a H. of D.	0,590	0,036	0,967	0,225	4,109	—2267,902	14,20	13	0,36
st.afw.	0,031	0,054	0,059	0,212	0,191				
Hald	0,6	0,0	1,04	0,15	4,168	—2269,125	16,51	13	0,22
D & B	0,552	0,025	0,90	0,2	3,8	—2269,795	15,39	12	0,22
Weichs.	0,622	0,0	1,30	1,6	3,44	—2325,245	153,51	10	$<10^{-5}$
Burrau	0,997	0,118	2,599	—0,312	9,323	—2421,240	326,25	9	$<10^{-5}$
1 Normale		0,114	2,736			—2431,101	321,23	12	$<10^{-5}$

aan de tweede. Als de verdelingen minder duidelijk gescheiden liggen blijft op deze manier een te groot indifferentiegebied over; men kan dan beter elke waarneming $\underset{\sim}{x} = x$ met kans $\varrho(x)$ aan de tweede en kans $1 - \varrho(x)$ aan de eerste component toewijzen.

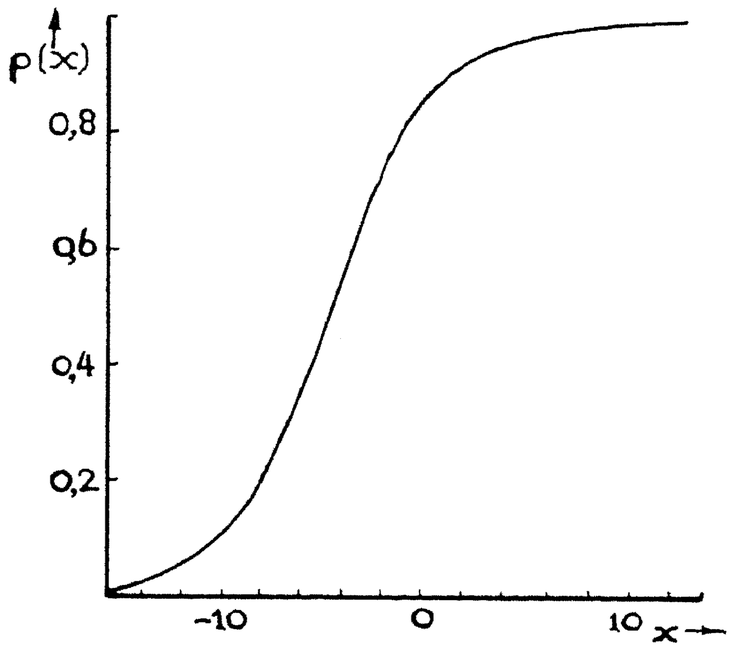

Fig. 1.

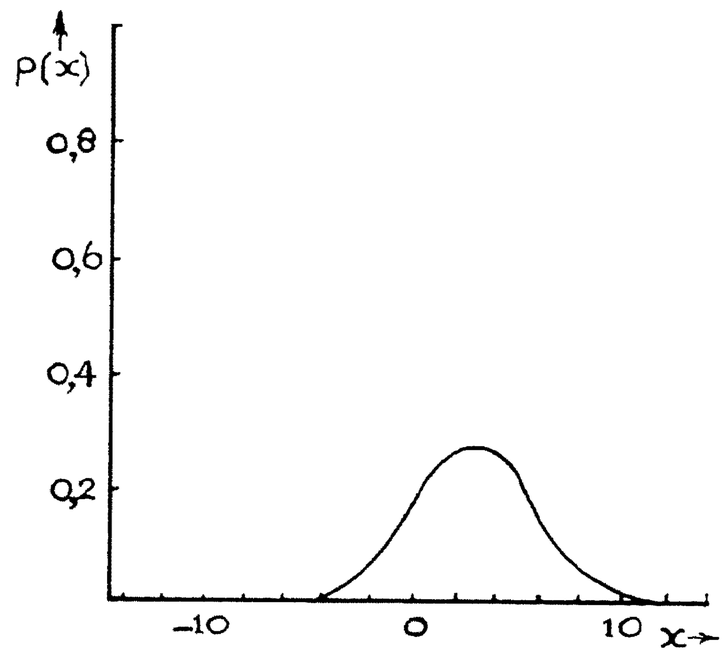

Fig. 2.

Locale mengverhouding $p(x)$ voor de schattingen ML v/a H. of R. (fig. 1) en ML v/a B.I. (fig. 2).

Voorbeelden

De besproken methoden werden toegepast op vijf voorbeelden van mengsels van normale verdelingen. Zoals uit tabel 1 blijkt, verschillen de voorbeelden in aantal waarnemingen, in verhouding van de standaardafwijkingen en in de verhouding van het verschil der verwachtingen tot het harmonisch gemiddelde van de varianties. Om eventuele beïnvloeding bij de grafische methoden te voorkomen, geschiedde de keuze van de parametervectoren en de productie van de aselecte steekproeven buiten mijn voorkennis. In tabel 1 vindt men de steekproefgegevens, waarbij de „waargenomen" aantallen soms vrij aanzienlijk afwijken van wat uit de onderliggende verdelingen kon worden verwacht (zie ook de χ^2-toetsen voor de regels „exact" in tabel 2). Dit geeft een goede gelegenheid om iets te constateren over de robuustheid van de diverse methoden tegen dit soort afwijkingen; vooral voor WEICHSELBERGER is het resultaat vrij somber. De extreme invloed van een paar uitbijters op momentenschatters blijkt bij het derde en vijfde voorbeeld.

Ter vergelijking van de aanpassing werden χ^2-toetsen uitgevoerd, waarbij het aantal vrijheidsgraden variëert met het aantal aangepaste parameters maar ook met het aantal klassen: steeds werden alleen die klassen samengevoegd die een verwacht aantal waarnemingen van minder dan vijf hadden. In de kolom log L van tabel 2 vindt men bovendien de logarithme van de aannemelijkheidsfunctie; de verschillen in log L tussen de ontmengingen van één voorbeeld vormen een criterium dat vrij goed met de χ^2-toets blijkt overeen te stemmen.

In tabel 2 geeft „exact" de onderliggende mengverdeling aan en „1 Normale" het resultaat van het aanpassen van één normale verdeling. Verder duidt „=" de hypothese van gelijke varianties aan en „$\frac{1}{2}$" de hypothese van mengverhouding $\frac{1}{2}$. Bij de meest aannemelijke schatters „ML" is vermeld van welke beginschattingen werd uitgegaan, terwijl hier tevens de geschatte standaardafwijkingen zijn vermeld. Als een methode twee verschillende uitkomsten suggereerde, zijn die onderscheiden door „I" en „II". Wanneer een besproken methode bij één of meer voorbeelden niet is vermeld, betekent dit dat er geen of onzinnige resultaten werden geproduceerd [1]).

De figuren 3 en 6 t/m 8 (door de rekenmachine vervaardigd) geven telkens de onderliggende verdeling (curve 1), de meest aannemelijke grafische schatting (curve 2) en de meest aannemelijke schatting (curve 3). Voor voorbeeld 2 zijn daarentegen zes curven geschetst, in de figuren 4 en 5.

---

[1]) Dit geldt niet voor het ontbreken van Burrau = en Rao = bij voorbeeld 4.

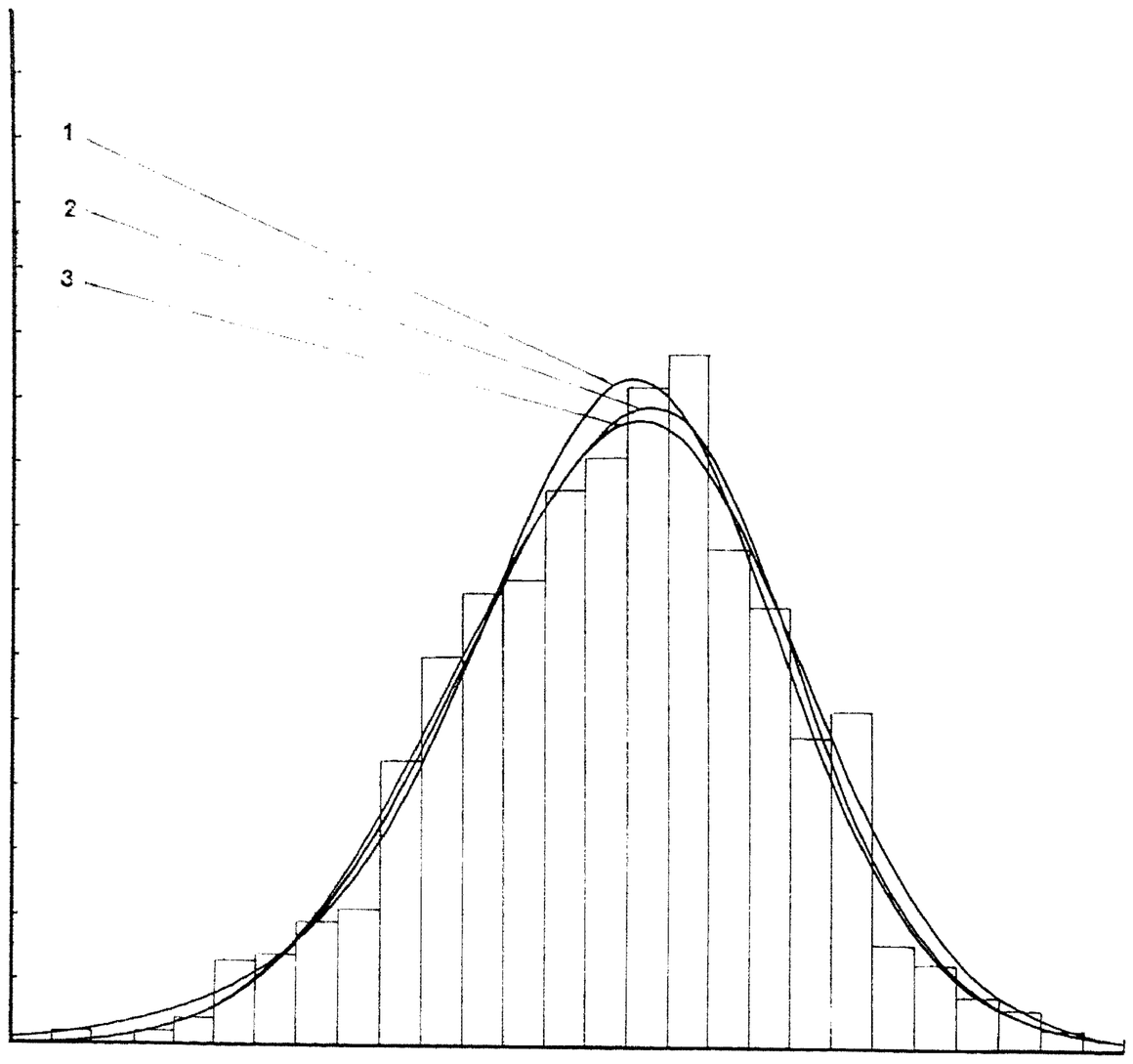

Fig. 3. VOORBEELD 1: exacte verdeling (curve 1), beste grafische schatting (Hald, curve 2) en meest aannemelijke schatting (curve 3), vergeleken met het histogram van de waarnemingen.

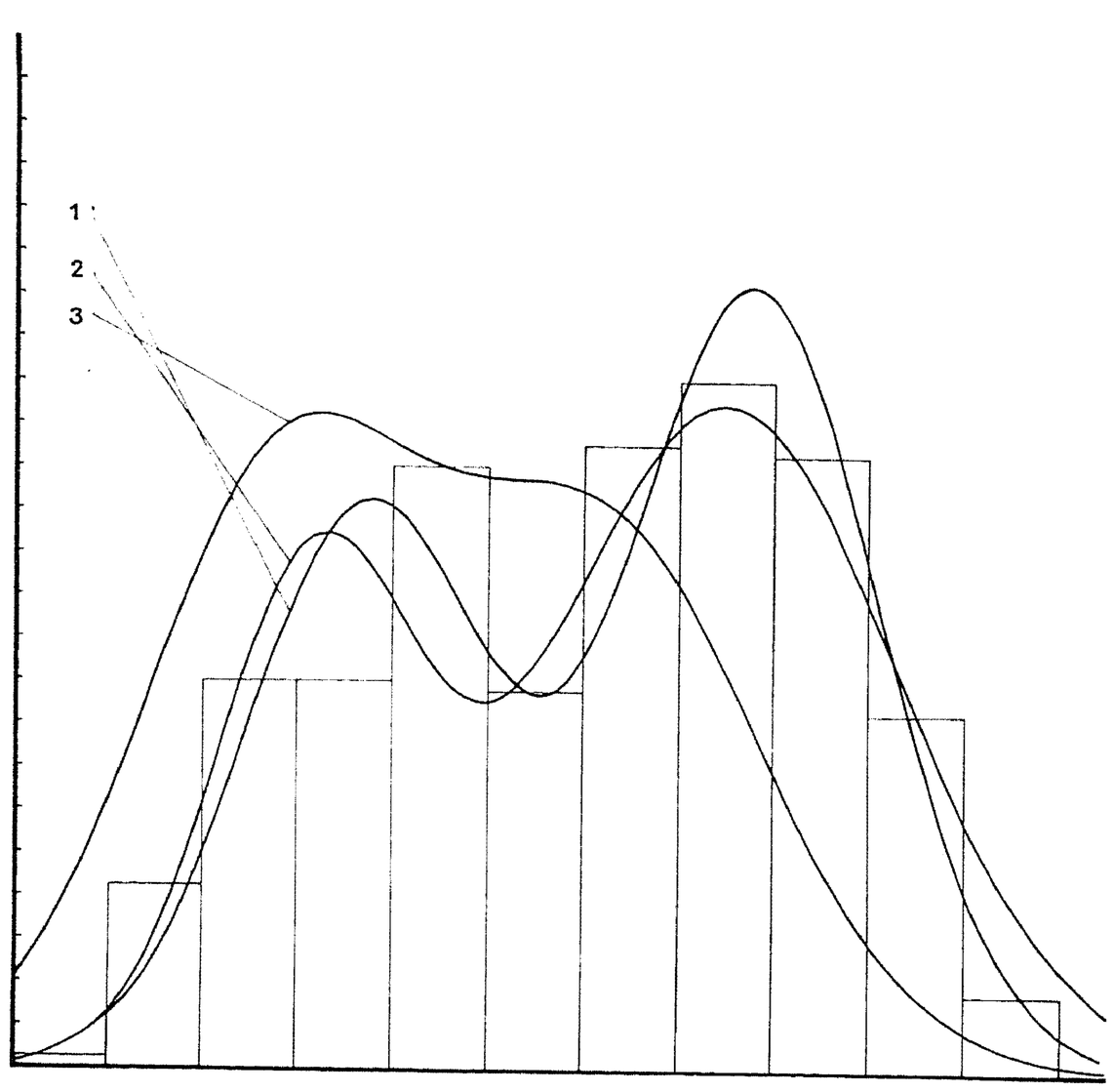

Fig. 4. VOORBEELD 2: exacte verdeling (curve 1), beste grafische schatting (Hald, curve 2) en schatting Weichselberger I (curve 3), vergeleken met het histogram van de waarnemingen. Zie ook fig. 5.

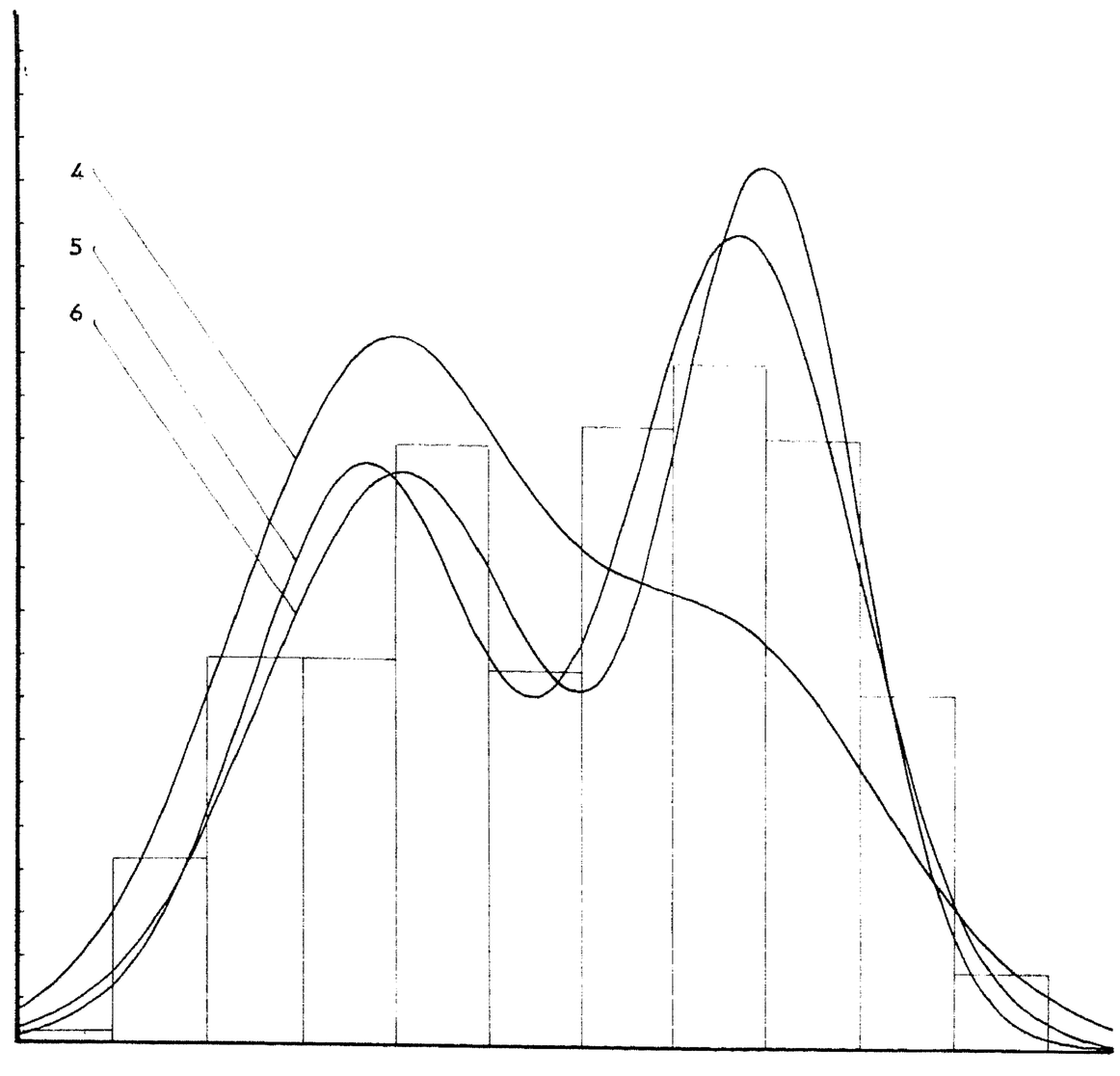

Fig. 5. VOORBEELD 2 vervolg: schatting Weichselberger II (curve 4), schatting Rao = (curve 5) en schatting Sittig $\frac{1}{2}$ (curve 6), vergeleken met het histogram van de waarnemingen. Zie ook fig. 4.

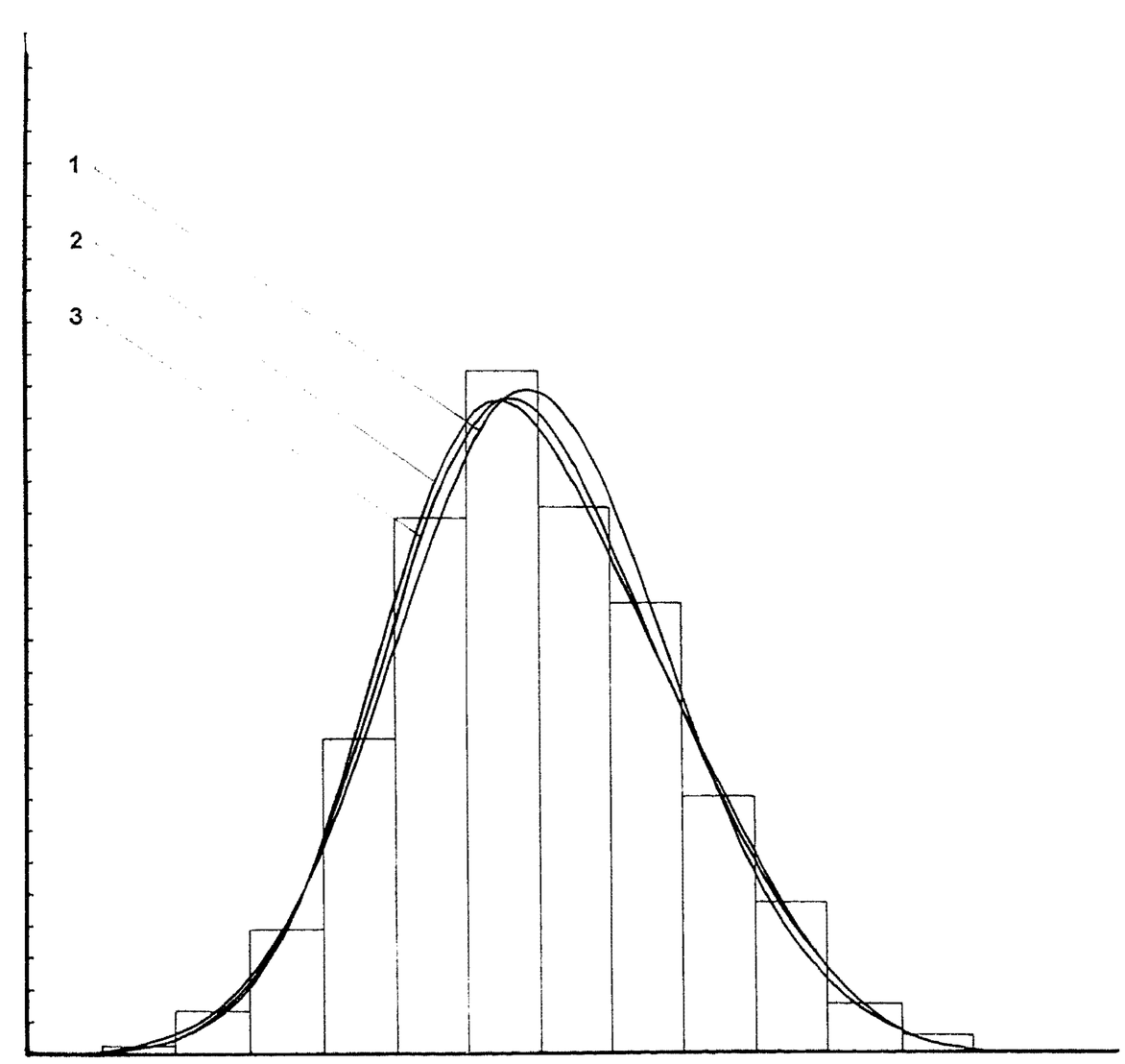

Fig. 6. VOORBEELD 3: exacte verdeling (curve 1), beste grafische schatting (Hald, curve 2) en meest aannemelijke schatting (curve 3), vergeleken met het histogram van de waarnemingen.

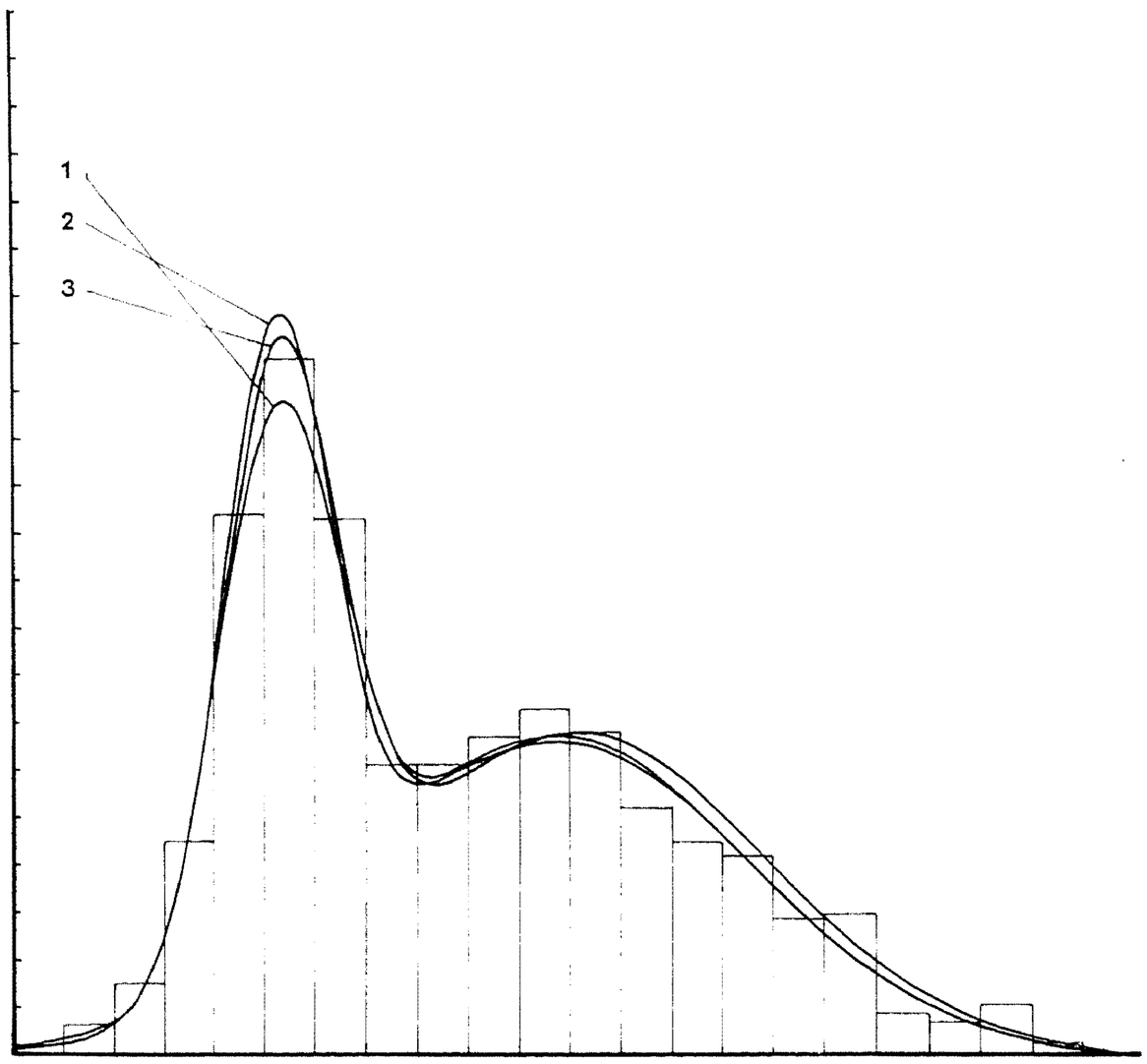

Fig. 7. VOORBEELD 4: exacte verdeling (curve 1), beste grafische schatting (Daeves en Beckel, curve 2) en meest aannemelijke schatting (curve 3), vergeleken met het histogram van de waarnemingen.

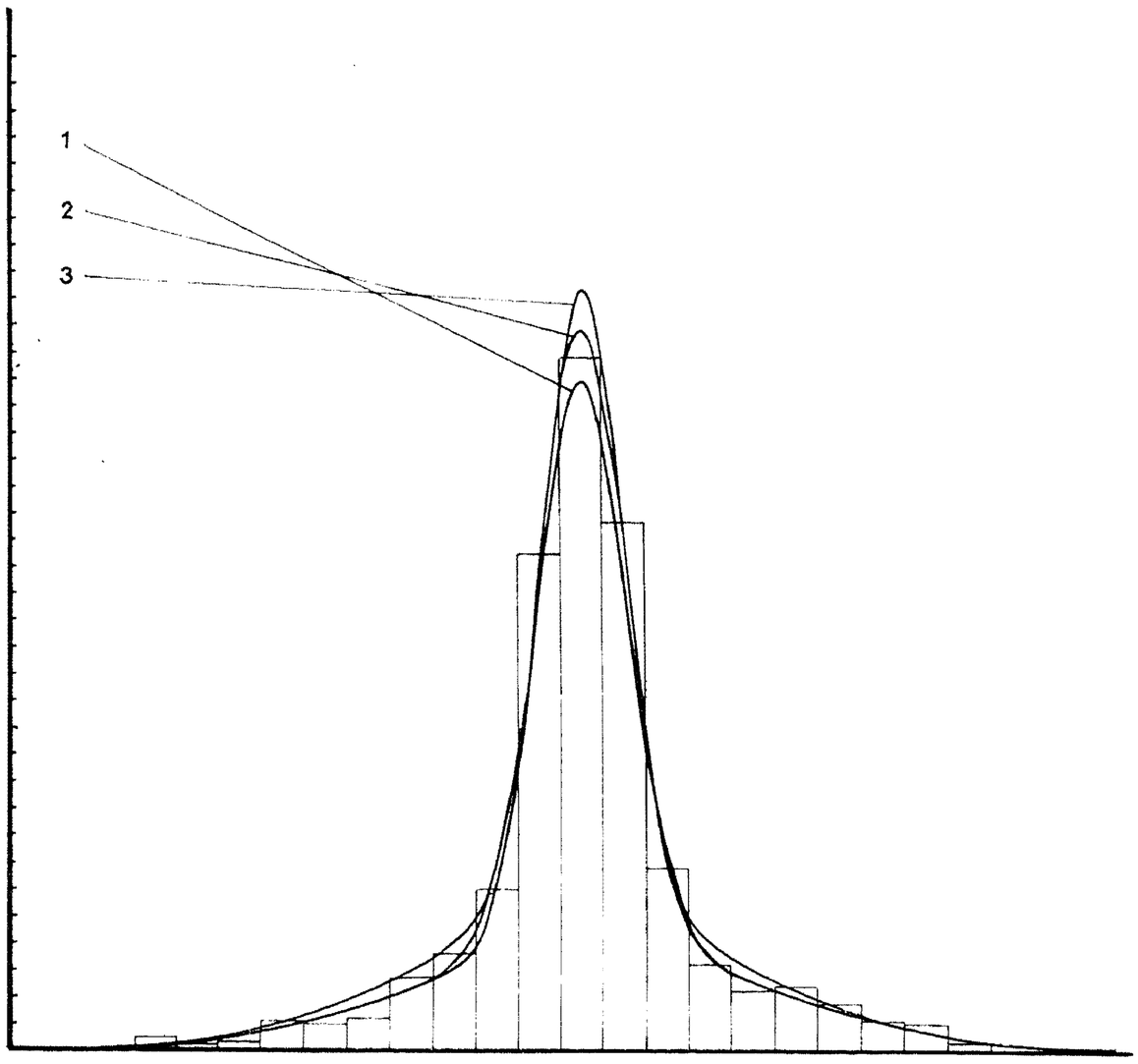

Fig. 8. VOORBEELD 5: exacte verdeling (curve 1), beste grafische schatting (Hald, curve 2) en meest aannemelijke schatting (curve 3), vergeleken met het histogram van de waarnemingen.

Dankwoord

Dit rapport zou niet, of veel later, zijn verschenen als ik niet de voortdurende medewerking had genoten van de technisch assistenten van het Mathematisch Centrum, die o.l.v. de heer VISSER veel rekenwerk hebben verricht, en van de heren ANTHONISSE en SCHWERDT, die geduldig bemiddelden tussen mij en de rekenmachine. Verder dank ik Dr. C. LEVERT en de Heer J. A. UYTERLINDEN, die mij attendeerden op enige publicaties. Laatstgenoemde heeft bovendien de tekeningen gecopiëerd.

Literatuur

[1] AGARD, J., Mélange de deux populations normales et étude de quelques fonctions $f(x, y)$ de variables normales x, y, Revue de Stat. Appl. 9 (1961), 53–70.

[2] VAN ALPHEN, G. N., Gesuperponeerde normale distributies en splitsing in elementaire curven, Stat. Neerl. 5 (1952), 47–53.

[3] BURRAU, C., The half-invariants of the sum of two typical laws of errors, with an application to the problem of dissecting a frequency curve into components, Skandin. Aktuarietidskr. 17 (1934), 1–6.

[4] CHARLIER, C. V. L., Researches into the theory of probability, Lunds Universitets Årsskrift, ny följd, afd. 2, 1 (1906), no. 5.

[5] COURT, A., Separating frequency distributions into two normal components, Science 110 (1949), 500–501.

[6] DAEVES, K. en BECKEL, A., Groszzahl-Methodik und Häufigkeitsanalyse, 2. Auflage, Weinheim Bergstr. (1958).

[7] ESSENWANGER, O., Neue Methode der Zerlegung von Häufigkeitsverteilungen in Gausz-sche Normalkurven und ihre Anwendung in der Meteorologie, Berichte des Deutschen Wetterdienstes No. 10 (1954).

[8] ESSENWANGER, O., Probleme der Häufigkeitsanalyse, Meteorologische Rundschau 7 (1954), 85–88.

[9] ESSENWANGER, O., Tafeln zur Häufigkeitszerlegung mit Anwendungsbeispielen, Berichte des Deutschen Wetterdienstes No. 39 (1957).

[10] HALD, A., Statistical theory with engineering applications, New York, (1952), 152–158.

[11] PEARSON, K., Contribution to the mathematical theory of evolution, Philos. Trans. of the Royal Society of London, Series A 185 (1894), 71–110; tevens in: Karl Pearson's early statistical papers, Cambridge 1948, 1–40.

[12] PRESTON, E. J., A graphical method for the analysis of statistical distributions into two normal components, Biometrika 40 (1953), 460–464.

[13] RAO, C. R., The utilization of multiple measurements in problems of biological classification, Journal of the Royal Statistical Society, Series B, 10 (1948), 159–203; het gedeelte over momentenschatters (p. 163–166) tevens in: C. R. Rao, Advanced statistical methods in biometric research, Wiley, New York 1952, 300–304.

[14] SITTIG, J., Superpositie van twee frequentieverdelingen, Stat. Neerl. 2 (1948), 206–227.

[15] STRÖMGREN, B., Tables and diagrams for dissecting a frequency curve into components by the half-invariant method, Skandin. Aktuarietidskr. 17 (1934), 7–54.

[16] TEICHER, H., Identifiability of finite mixtures, Ann. Math. Stat. 34 (1963), 1265–1269.

[17] WEICHSELBERGER, K., Über ein graphisches Verfahren zur Trennung von Mischverteilungen und zur Identifikation kupierter Normalverteilungen bei groszem Stichprobenumfang, Metrika 4 (1961), 178–229.