LES CAHIERS DE L'ANALYSE DES DONNÉES

P. CAZES

Note sur l'estimation de courbes bimodales

Les cahiers de l'analyse des données, tome 4, n° 3 (1979), p. 331-338

<http://www.numdam.org/item?id=CAD_1979__4_3_331_0>

Les Cahiers de l'Analyse des Données
Vol. IV - 1979 - n° 3 - p. 331-338

NOTE SUR L'ESTIMATION DE COURBES BIMODALES [COURBES BIMODALES]

par P. Cazes (1)

1 *Introduction - Rappels*

Cette note fait suite à l'article [PHOTOMULTIPLICATEUR](cf *Cahier*, Vol III n° 4 pp. 393-417) et donne les résultats de simulations effectuées pour estimer des courbes bimodales. Rappelons brièvement le problème et la manière dont on l'a résolu : étant donné une variable aléatoire Y, dont la loi est de la forme : $y(x) = \int_a^b g(x')\, k(x',x)dx'$, on cherche, $k(x',x)$ étant connu à estimer la fonction g (qui est en fait une loi de probabilité) à partir d'un histogramme de Y. Pour effectuer cette estimation, on découpe l'intervalle (a,b) de variation de g en p intervalles, et l'on est ramené, si J désigne l'ensemble des classes de l'histogramme de Y, et y_j la proportion d'observations tombant dans la j-ème classe de cet histogramme, au système d'équations :

$$\forall\, j \in J : y_j \approx \Sigma\{b_i k(i,j) \mid i = 1, p\}$$

système où l'on a posé : $b_i = g(c_i)\ \Delta x_i$, c_i étant le centre du i-ème intervalle du découpage adopté, intervalle de longueur Δx_i, et où $k(i,j)$ est la valeur de l'intégrale de $k(c_i, x)$ sur la j-ème classe de l'histogramme.

L'estimation des coefficients b_i (qui doivent être positifs pour avoir un sens) permet alors d'estimer les valeurs de la fonction g en p points d'où l'intérêt de prendre une valeur de p assez élevée pour avoir une estimation "fine" de g.

Pour estimer ces coefficients, on se place dans l'espace des r premiers facteurs de l'analyse des correspondances du tableau k_{IJ} des $k(i,j)$; le nombre r est déterminé de telle sorte que la projection de y_J* sur l'espace E_r des r premiers axes factoriels (dans R_J) soit à l'intérieur du convexe C_r engendré par les combinaisons linéaires à coefficients positifs et de somme inférieure ou égale à 1 des projections des

$$k_{iJ} = \{k(i,j) \mid j \in J\},\ (1 \le i \le p),$$

tandis que la projection de y_J dans l'espace E_{r+1} associé aux r+1 premiers axes factoriels, est à l'extérieur du convexe C_{r+1} associé.

* $y_J = \{y_j \mid j \in J\}$

(1) Maître-Assistant, laboratoire de statistique, Université P. & M. Curie

Posant $yy_J = \Sigma\{b_i\ k_{iJ} \mid i = 1, p\}$ les coefficients b_i sont alors déterminés de telle sorte que le carré de la norme $\|y_J - yy_J\|^2$ soit minimum, avec les contraintes : $yy_J \in E_r$; $b_1 \geq 0, \ldots, b_p \geq 0$, R_J étant muni de la métrique du χ^2 de centre k_J/k, k_J désignant la marge sur J de k_{IJ} et k le total des éléments du tableau k_{IJ}.

Après avoir estimé de cette manière la statistique de multiplication du premier étage d'un photomultiplicateur à dynodes, nous avions montré, par simulation que la méthode d'estimation utilisée permettait de reconstituer des lois g unimodales (lois normales ou lois gamma). Par contre pour certaines lois g bimodales, mélange de deux lois normales imbriquées, du fait d'un certain lissage, nous ne retrouvions la bimodalité que sous certaines conditions, évoquées en remarque à la fin de l'article [PHOTOMULTIPLICATEUR], d'où l'étude présentée ici, étude essentiellement expérimentale.

2 *Résultats sur l'estimation de courbes bimodales*

Rappelons que se donnant une loi g (ici bimodale), on en déduit (un découpage de l'intervalle (a,b) de variation de g ayant été effectué *) les coefficients b_i, et donc l'histogramme y_J associé :

$$y_J = \Sigma\{b_i\ k_{iJ} \mid i = 1, p\}$$

On considère aussi la loi y'_J déduite de y_J par une perturbation aléatoire :

$$\forall\ j \in J : y'_j = y_j + \sqrt{y_j}\ e_j$$

où les e_j sont des erreurs normales centrées indépendantes, et de même variance ε^2. Plusieurs valeurs de ε ont été considérées entre $0 (y'_J = y_J)$ et 10^{-2}. Les résultats qui diffèrent peu en général quand ε varie dans la plage précédente, sont donnés dans le cas où $\varepsilon = 5 \times 10^{-3}$.

Les courbes g que nous avons essayé de reconstituer par la méthode rappelée au § 1 sont des mélanges de lois normales de la forme :

$$p_1\ LG(m_1, \sigma_1) + p_2\ LG(m_2, \sigma_2)$$

$LG(m_i, \sigma_i)$ désignant la loi de Laplace-Gauss de moyenne m_i, et d'écart-type σ_i, et p_i la proportion de cette loi ($1 \leq i \leq 2$; $p_2 = 1 - p_1$).

Dans tous les essais effectués, essais numérotés de 1 à 10, et reportés sur le tableau 1, on a choisi les valeurs de m_1 et de m_2 de telle sorte que $m_1 + m_2 = 15$, ce qui donne une moyenne égale à 7,5 quand les proportions p_1 et p_2 sont égales ($p_1 = p_2 = 0{,}5$), ce qui était le cas sauf pour un essai. On a également choisi (sauf pour deux essais) des écarts-type σ_1 et σ_2 égaux. On a retenu trois valeurs pour m_1 (et donc pour $m_2 = 15 - m_1$) : 4, 5 et 6, ainsi que trois valeurs d'écart-type : 1, 1,5 et 2.

---

* *Nous avons adopté le même découpage en 31 classes que précédemment (cf [PHOTOMULTIPLICATEUR] § 3.3 in fine), des découpages voisins ayant fourni, comme il fallait s'y attendre, mais comme on l'a vérifié, des résultats similaires à ceux présentés ici. Ce découpage correspond à un pas de 0,5 entre 0,5 et 14, puis un pas de 1 entre 14 et 17.*

On n'a représenté sur les figures 1 à 6 que sept des dix cas étudiés ici, ces sept cas résumant bien l'ensemble des situations rencontrées, tandis que sur le tableau 1, on a reporté pour chacune des dix estimations le nombre r de facteurs conservés pour réaliser cette estimation, ainsi que les valeurs des statistiques d'ajustement T^2 et U^2 suivantes :

$$T^2 = \Sigma\{(b_i - \hat{b}_i)^2 | i = 1, p\}$$

$$U^2 = \Sigma\{(b_i - \hat{b}_i)^2/(b_i + \hat{b}_i) | i = 1, p\}$$

$\hat{b}_i$ étant l'estimation de b_i obtenue par la méthode de régression utilisée.

essai	m_1	σ_1	p_1	m_2	σ_2	p_2	r	10^4T^2	10^3U^2
1	4	1	0,5	11	1	0,5	7	66	128
2	4	1,5	0,5	11	1,5	0,5	7	14	39
3	4	2	0,5	11	2	0,5	5	12	23
4	4	1	0,75	11	1	0,25	7	46	94
5	5	1,5	0,5	10	1,5	0,5	7	20	50
6	5	2	0,5	10	2	0,5	7	4,7	27
7	5	2	0,5	10	1,5	0,5	7	12	37
8	5	1,5	0,5	10	2	0,5	6	6,7	33
9	6	1	0,5	9	1	0,5	6	50	84
10 *	6	1	0,5	9	1	0,5	11	17	73

Tableau n° 1 : Estimation d'un mélange de deux lois normales $p_1LG(m_1, \sigma_1) + p_2LG(m_2, \sigma_2)$, *perturbé par une erreur normale d'écart-type* $\varepsilon = 5 \times 10^{-3}$.

r : nombre de facteurs conservés pour l'estimation.

T^2, U^2 *: statistiques permettant de juger de l'adéquation de l'estimation obtenue. Dans le cas n° 10, (mélange identique au cas n° 9) l'estimation est faite à partir d'une combinaison linéaire de lois explicatives d'écart-type deux fois plus petit que dans les autres essais.*

Dans tous les essais, sauf dans le neuvième ($m_1 = 6$, $m_2 = 9$, $\sigma_1 = \sigma_2 = 1$, $p_1 = p_2 = 0,5$) sur lequel nous reviendrons, la bimodalité a été retrouvée, mais on a constaté un certain lissage, lissage d'autant plus important qu'on a des pics plus étroits ($\sigma_1 = \sigma_2 = 1$; cf essais n° 1 et 4 et figures 1 et 2); on remarque que pour un écart-type égal à 2, on retrouve bien la bimodalité quoiqu'elle soit très peu accentuée (cf essais 3 et 6, et la figure 4) la reconstitution étant excellente et meilleure que dans le cas où $\sigma_1 = \sigma_2 = 1,5$ (cf essais 2 et 5 et figure 3).

On peut noter que la forme des courbes est bien reconnue dans le cas où $p_1 \neq p_2$ (cas n° 4 où $p_1 = 0,75$, $p_2 = 0,25$) mais avec un lissage assez important des pics puisque $\sigma_1 = \sigma_2 = 1$ (cf figure 2), ainsi que dans le cas où $\sigma_1 \neq \sigma_2$ (cf essais 7 et 8 et figure 5).

Par contre, dans le cas n° 9, la proximité des deux pics ($m_1 = 6$, $m_2 = 9$, $m_2 - m_1 = 3\sigma_1 = 3\sigma_2$) et le lissage de chacun d'eux entraînent qu'on obtient une estimation unimodale de la courbe testée. En fait, quand $\varepsilon = 0$ (auquel cas on est sûr de reconstituer la courbe exactement, si on garde un nombre suffisant de facteurs), la bimodalité n'est reconnue qu'à partir du 13° facteur ; or pour des valeurs de ε non nulles, comprises entre 10^{-2} et 10^{-4}, la valeur maximale de r pour laquelle la projection de y'_J dans l'espace E_r des r premiers axes factoriels est à l'intérieur du convexe C_r associé (cf § 1) varie entre 6 et 9 suivant la valeur de ε. Il en résulte que l'estimation obtenue pour la courbe g est unimodale. Pour retrouver la bimodalité, il faut prendre une perturbation de variance plus faible, de façon à ce que la projection de y'_J dans l'espace des 13 premiers axes factoriels soit à l'intérieur du convexe C_{13}, ce qui est réalisé comme on l'a signalé dans l'article [PHOTOMULTIPLICATEUR], pour des valeurs de ε inférieures ou égales à 10^{-5}.

Une autre manière pour reconnaître la bimodalité est de prendre des lois explicatives k_{iJ} moins dispersées ; à la limite, si on a des lois de Dirac, les k_{iJ} n'interfèrent plus et sont orthogonaux, ce qui permet une meilleure estimation que dans le cas de lois k_{iJ} interférant (pour les 31 lois k_{iJ} associées au découpage adopté, on a avec les quatre premiers facteurs de l'analyse factorielle plus de 99% de l'inertie totale, alors que dans le cas de lois de Dirac, toutes les valeurs propres étant égales à 1, on aurait 4/30 = 13,33% de l'inertie avec quatre facteurs).

A toute loi k_{iJ}, correspondant au découpage adopté, loi de moyenne μ_i et d'écart-type s_i, on a donc associé une loi h_{iJ} de même moyenne et d'écart-type s_i/A, A^2 étant un facteur de réduction de la variance. Les lois k_{iJ} ayant la forme de lois gamma (cf [PHOTOMULTIPLICATEUR] § 2.1 pour la définition et le calcul de ces lois), on a adopté pour h_{iJ} la loi gamma de moyenne μ_i et d'écart-type s_i/A (i. e. la loi d'une variable aléatoire z telle que $az \in \gamma_t$, avec $E(z) = t/a = \mu_i$; $Var\ z = t/a^2 = s_i^2/A^2$, d'où l'on déduit facilement t et a en fonction de μ_i, s_i et A).

Prenant ainsi une combinaison linéaire des h_{iJ} (et non des k_{iJ}) la bimodalité a été retrouvée dans ce cas (cas numéroté 10 dans le tableau 1, pour le différencier du cas n° 9, ces deux essais correspondant au même mélange, i.e. aux mêmes coefficients b_i, le premier étant relatif à la combinaison linéaire des h_{iJ} et le second à celle des k_{iJ}) pour un facteur de réduction A supérieur ou égal à 2, et l'estimation correspondante (pour A = 2) est reportée sur la figure 6, cette estimation ayant été effectuée avec 11 facteurs. On constate encore, bien que la modalité ait été retrouvée, un certain lissage des pics, mais moins important qu'auparavant.

Remarque : Comme on l'avait déjà constaté dans l'article [PHOTO - MULTIPLICATEUR] § 4.3, on a encore ici dans la plupart des essais effectués de légères anomalies (d'amplitude relativement faible) au début et à la fin des histogrammes estimés, où les estimations sont donc peu précises, ce qui se comprend aisément puisque dans tous les essais les lois g(x) testées étaient nulles pour les faibles où les fortes valeurs de x.

En conclusion, avec la méthode de régression proposée, on peut reconnaître avec les lois k_{iJ} considérées des bimodalités même peu accentuées, à condition que les pics ne soient pas trop aigus. Si les pics sont aigus et bien séparés, ils sont reconnus, mais avec un lissage assez fort. Si les pics sont aigus et interfèrent, on risque du fait du lissage de de ne pouvoir détecter la bimodalité.

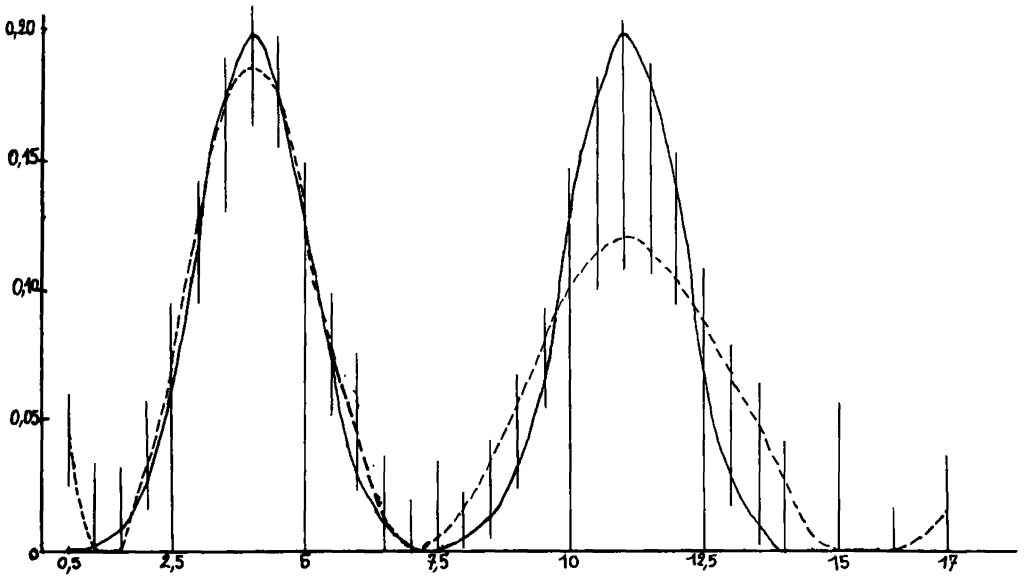

Figure 1 (cas n° 1) : Estimation d'un mélange g de deux lois normales ($m_1 = 4$, $m_2 = 11$) de mêmes proportions et de mêmes écarts-types ($\sigma_1 = \sigma_2 = 1$).

Pour tracer la courbe associée à g et son estimation, on a joint de façon continue les valeurs obtenues aux extrémités du découpage adopté, extrémités caractérisées par les traits verticaux sur le graphique.

——— : *loi g*

- - - - - : *estimation de g*

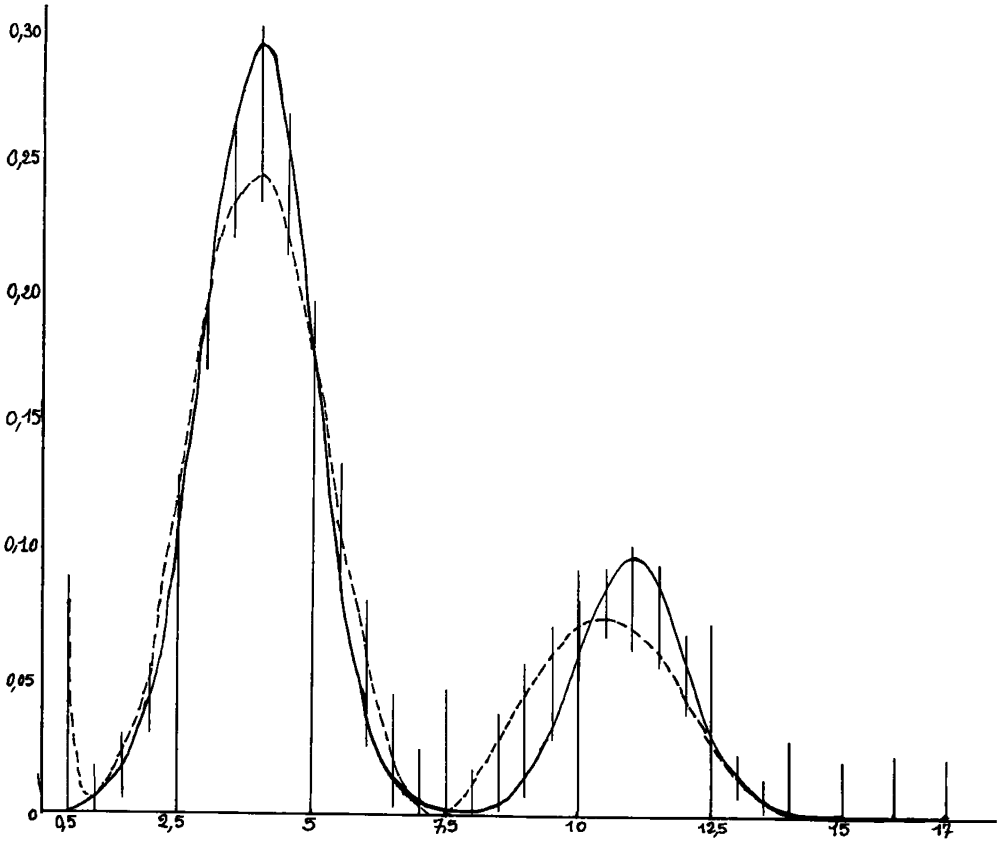

Figure 2 (cas n° 4) : Estimation d'un mélange g de lois normales ($m_1 = 4$, $m_2 = 11$) de mêmes écarts-types ($\sigma_1 = \sigma_2 = 1$) et en proportions différentes ($p_1 = 0,75$; $p_2 = 0,25$).

——— : *loi g*

- - - - - : *estimation de g*

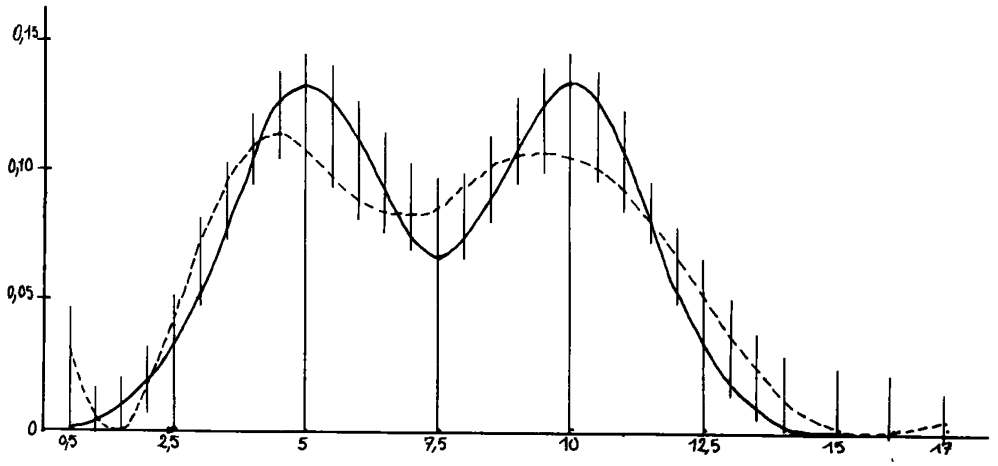

Figure 3 (cas n° 5) : $\sigma_1 = \sigma_2 = 1,5$

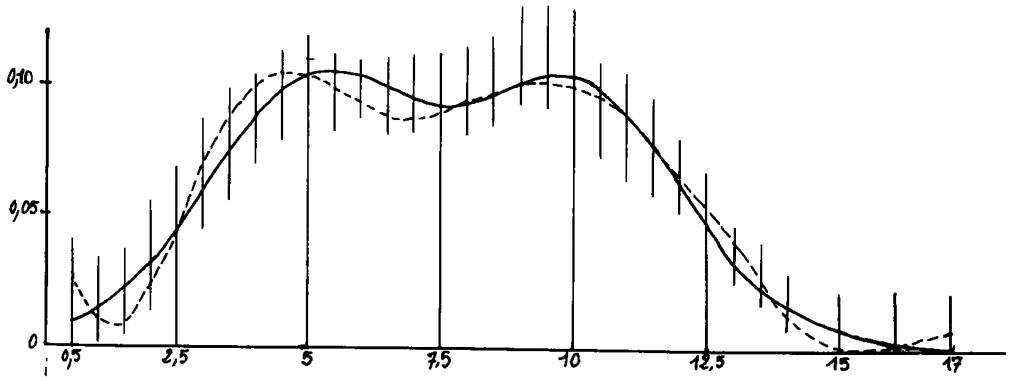

Figure 4 (cas n° 6) : $\sigma_1 = \sigma_2 = 2$

Figures 3 et 4 : Estimation d'un mélange g de deux lois normales ($m_1 = 5$, $m_2 = 10$) de mêmes écarts-types et de mêmes proportions.

——— : *loi g*

- - - - - : *estimation de g*

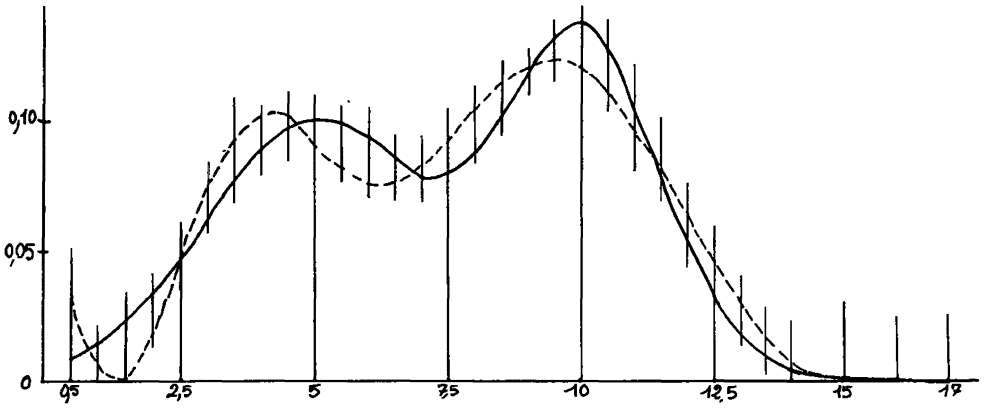

Figure n° 5 (cas n° 7) : Estimation d'un mélange g de deux lois normales ($m_1 = 5$, $m_2 = 10$) de mêmes proportions, et ayant des écarts-types différents ($\sigma_1 = 2$; $\sigma_2 = 1,5$).

——— : *loi g*

- - - - - : *estimation de g*

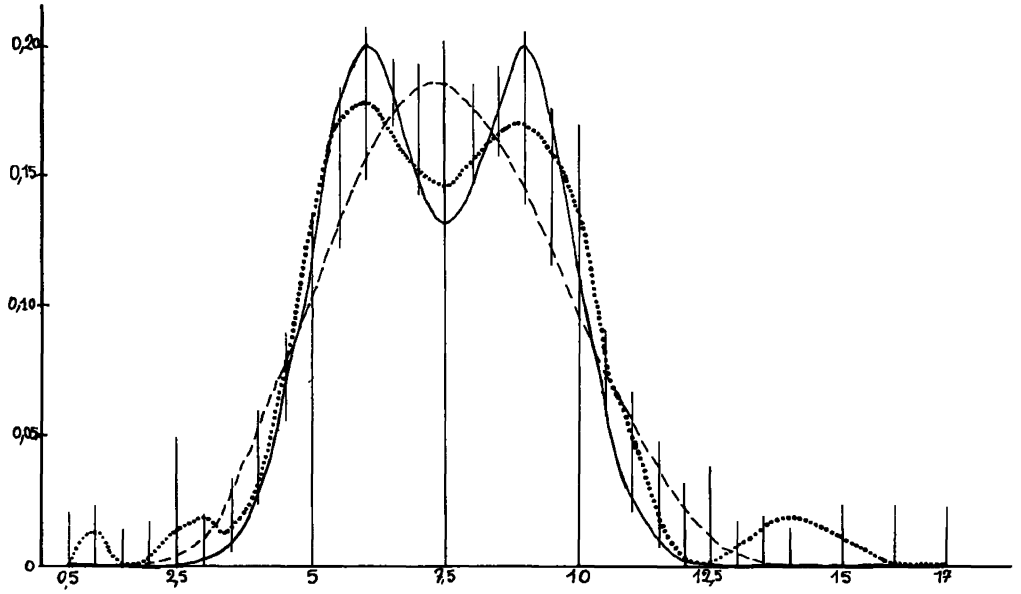

Figure 6 (cas n^{os} 9 et 10) : Estimation d'un mélange de deux lois normales imbriquées, de mêmes proportions et de mêmes écarts-types ($m_1 = 6$; $m_2 = 9$; $\sigma_1 = \sigma_2 = 1$).

——— : *loi g*

- - - - - : *estimation de g dans le cas n° 9.*

. . . . . : *estimation de g dans le cas n° 10 où les lois explicatives sont deux fois moins dispersées que dans le cas n° 9.*