

ROBERT PAOLI

Réflexions sur la détermination des fonctions de distribution

Journal de la société statistique de Paris, tome 120, n° 4 (1979), p. 244-269

http://www.numdam.org/item?id=JSFS_1979__120_4_244_0

© Société de statistique de Paris, 1979, tous droits réservés.

L'accès aux archives de la revue « Journal de la société statistique de Paris » (<http://publications-sfds.math.cnrs.fr/index.php/J-SFdS>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

RÉFLEXIONS SUR LA DÉTERMINATION DES FONCTIONS DE DISTRIBUTION

Robert PAOLI

*Ancien élève de l'École polytechnique
Administrateur de « L'Entreprise Industrielle »*

En présence d'une distribution empirique des fréquences d'une variable aléatoire, est-il véritablement justifié d'avoir recours à la loi de Gauss ayant même moyenne et même écart type pour définir la loi de répartition de la population mère? A cette méthode, dite du maximum de vraisemblance, on peut en substituer une autre qui minimise l'ensemble des écarts entre les fréquences observées et les fréquences résultant de la loi cherchée. Il est d'ailleurs aisé de montrer en quoi diffèrent les deux démarches et comment elles se situent par rapport au critère de Pearson.

Au surplus la référence à la loi de Gauss peut n'être pas justifiée et il est des cas où l'utilisation d'une généralisation de la Loi de Poisson est susceptible de donner de bons résultats. Un extrait des tables correspondantes est donné en annexe.

When there is an empirical distribution of random variable frequencies, is it really necessary to use the Gauss law with the same average and standard deviation to define the parent population distribution law? To this method, so-called maximum likelihood, it is possible to substitute another one that minimizes the total deviations between the observed frequencies and the frequencies resulting from the desired law. It is besides easy to show in what the two procedures are different and how they can be situated in relation with the Pearson criterium.

Moreover, the reference to the Gauss law may be not justified and there are some cases where the use of a generalization of the Poisson law may yield good results. An abstract of the corresponding tables is given in annex.

Der Verfasser stellt sich die Frage, ob im Falle einer zufälligen Verteilung der Häufigkeiten einer unbestimmten Variablen es nötig ist die Formel von Gauss anzuwenden im Falle der gleichen Durchschnitte und der gleichen Abweichung um das Gesetz der Verteilung der ursprünglichen Grundzahlen zu definieren? Im übrigen kann die Bezugnahme auf das Gesetz von Gauss nicht berechtigt sein und es gibt Fälle, wo die Verwendung einer Verallgemeinerung der Formel von Poisson sehr gute Resultate gibt. In der Beilage ist ein Auszug der entsprechenden Tafeln gegeben.

Bien que la loi normale ou de Gauss n'ait plus à l'heure actuelle la même faveur auprès des théoriciens de la statistique, c'est à elle cependant qu'ont généralement recours les praticiens devant une répartition empirique lorsqu'elle leur paraît à peu près symétrique. L'ajustement se fait alors par identification des moyennes et des écarts-types de la loi normale d'une part, de la répartition empirique d'autre part. Voilà une méthode simple qui permet de disposer, avec facilité des propriétés de la loi de Gauss dans l'analyse de la répartition étudiée. On sait que la justification de cette méthode réside dans le fait que la loi à laquelle elle conduit est celle qui est la plus vraisemblable. On sait aussi que, bien que la plus vraisemblable, ce n'est pas la plus probable.

L'expérience nous a amené à approfondir les motifs du choix de la méthode et à proposer finalement une solution différente de la précédente.

1. Le problème s'énonce ainsi : « En présence d'un échantillon de taille n extrait d'une population mère inconnue, comment déterminer au mieux les caractéristiques (moyenne écart-type) de cette dernière? »

Reprenons le calcul traditionnel qui a pour objet de caractériser la population mère d'où est extrait l'échantillon de taille n .

Soient $x_1 . x_2 . . . x_n$, les valeurs de la variable aléatoire composant l'échantillon et soit $\varphi(x)$ la densité de probabilité cherchée de la population-mère. La probabilité de réalisation de l'échantillon est :

$$P = \varphi(x_1) . \varphi(x_2) . . . \varphi(x_n) dx^n$$

Nous voulons déterminer $\varphi(x)$ de telle manière que la probabilité d'avoir obtenu l'échantillon en question soit maximum. Autrement dit, étant donné une forme de densité de probabilité $\varphi(x)$ dépendant d'un ou plusieurs paramètres, α , etc., nous voulons déterminer la valeur de ces derniers pour que P soit maximum, ce qui conduit à l'annulation des dérivées

partielles de P par rapport à ces paramètres : $\frac{\partial P}{\partial \alpha} = 0$, etc. ou ce qui revient au même

$$\frac{\frac{\partial P}{\partial \alpha}}{P} = 0, \quad \text{etc.}$$

En désignant par φ_i la valeur de la fonction $\varphi(x)$ pour $x = x_i$ et par $\left(\frac{\partial \varphi}{\partial \alpha}\right)_i$ la valeur de la dérivée partielle de $\varphi(x)$ par rapport à α pour $x = x_i$, il vient :

$$\sum_1^n \frac{\left(\frac{\partial \varphi}{\partial \alpha}\right)_i}{\varphi_i} = 0, \quad \text{etc.}$$

Jusqu'ici le raisonnement reste extrêmement général et l'on a toujours par annulation des dérivées partielles le nombre d'équations exactement nécessaire à la détermination des paramètres dont dépend la fonction $\varphi(x)$.

Remarquons que si $\varphi(x)$ dépend d'un certain nombre de paramètres α , etc. la moyenne μ de $\varphi(x)$ ainsi que son écart-type σ sont des fonctions desdits paramètres de sorte que lorsqu'on les a déterminés au moyen du système d'équations ci dessus, on peut aussi calculer μ et σ .

Nous n'avons repris ce raisonnement bien connu que pour montrer son degré de généralité et mieux permettre de le comparer à ceux qui vont suivre.

Nous allons maintenant poursuivre en le particularisant à la recherche d'une loi normale :

$$\varphi(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Les équations ci-dessus contiennent maintenant directement μ et σ et conduisent aux résultats connus $\mu = m$, $\sigma = s$ (1). Ainsi se trouve justifiée la règle qui consiste à identifier respectivement les moyennes et les écarts types. D'où également les commentaires concernant la valeur de σ qui selon cette méthode s'avère n'être pas la valeur la plus probable de

l'écart-type de la distribution empirique, laquelle est, comme chacun sait, $s \sqrt{\frac{n}{n-1}}$ et non pas s .

Nous avons raisonné sur les n valeurs de la variable $x, x_2 \dots x_n$ constituant l'échantillon; mais le plus souvent ces valeurs sont groupées en classes. Soit q le nombre de ces

classes et Δx leur intervalle. Dans l'intervalle $\left(x_i - \frac{\Delta x}{2}, x_i + \frac{\Delta x}{2}\right)$

$$\begin{aligned} \text{il y aura } f_i \text{ valeurs de la variable et l'on aura : } & \sum_{i=1}^{i=q} f_i = n \\ & \sum_{i=1}^{i=q} f_i x_i = n m \\ & \sum_{i=1}^{i=q} f_i (x_i - m)^2 = n s^2 \end{aligned}$$

Il n'est pas inutile de reprendre le raisonnement sur cette forme de répartition. La

loi de probabilité de la population-mère définit pour l'intervalle $\left(x_i - \frac{\Delta x}{2}, x_i + \frac{\Delta x}{2}\right)$

une probabilité de réalisation $p_i \Delta x$. Donc l'échantillon de taille n a une probabilité de réalisation

$$P = p_1^{f_1} \cdot p_2^{f_2} \dots p_q^{f_q} \Delta x^n$$

ce qui conduit par annulation des dérivées partielles à

$$\sum_{i=1}^{i=q} \frac{f_i}{p_i} \frac{\partial p_i}{\partial \alpha} = 0, \text{ etc.}$$

$$1. \text{ On a en effet } \frac{1}{\varphi(x)} \frac{\partial \varphi(x)}{\partial \mu} = \frac{x - \mu}{2 \sigma^2} \quad \text{d'où } \sum x_i = n \mu$$

$$\text{et de même } \frac{1}{\varphi(x)} \frac{\partial \varphi(x)}{\partial \sigma} = -\frac{1}{\sigma} + \frac{(x - \mu)^2}{\sigma^3} \quad \text{d'où } \sum (x_i - \mu)^2 = n \sigma^2$$

Nous supposerons maintenant que nous cherchons une fonction $\varphi(x)$ où interviennent directement les paramètres μ (moyenne de la population-mère) et σ (écart-type de la population-mère). Le raisonnement se continuera en écrivant

$$\left\{ \begin{array}{l} \sum_{i=1}^{i=q} \frac{f_i}{p_i} \frac{\partial p_i}{\partial \mu} = 0 \\ \sum_{i=1}^{i=q} \frac{f_i}{p_i} \frac{\partial p_i}{\partial \sigma} = 0 \end{array} \right. \quad (1)$$

On remarque que :

$$p_i \Delta x = \int_{x_i - \frac{\Delta x}{2}}^{x_i + \frac{\Delta x}{2}} \varphi(x) dx$$

et par conséquent, que :

$$\frac{\partial p_i}{\partial \mu} \Delta x = \int_{x_i - \frac{\Delta x}{2}}^{x_i + \frac{\Delta x}{2}} \frac{\partial \varphi(x)}{\partial \mu} dx$$

(de même pour $\frac{\partial p_i}{\partial \sigma}$)

En fait, on sait par définition de la moyenne, qu'il existe dans l'intervalle considéré une valeur x_i de la variable qui permet d'écrire

$$p_i = \varphi(x_i')$$

Il existe aussi une valeur x_i'' telle que

$$\frac{\partial p_i}{\partial \mu} = \frac{\partial \varphi(x_i'')}{\partial \mu}$$

(de même pour $\frac{\partial p_i}{\partial \sigma}$)

Δx étant petit, on peut accepter de remplacer x_i' et x_i'' par x_i et d'écrire, avec l'approximation que cela comporte :

$$\left\{ \begin{array}{l} \sum_{i=1}^{i=q} \frac{f_i}{\varphi(x_i)} \frac{\partial \varphi(x_i)}{\partial \mu} = 0 \\ \sum_{i=1}^{i=q} \frac{f_i}{\varphi(x_i)} \frac{\partial \varphi(x_i)}{\partial \sigma} = 0 \end{array} \right.$$

Cette approximation simplifie le système en ce sens que $\varphi(x)$ intervient cette fois directement et non plus par une série de quadratures ⁽¹⁾.

Si, comme il y a un instant, on cherche une loi normale, on retrouve :

$$\sum_{i=1}^{i=g} f_i (x_i - \mu) = 0$$

$$\sum_{i=1}^{i=g} f_i \left[\frac{-1}{\sigma} + \frac{(x_i - \mu)^2}{\sigma^3} \right] = 0$$

et les résultats connus $\mu = m$, $\sigma = s$
et la justification de la même règle.

Nous ferons une autre remarque : les dérivées logarithmiques prises par rapport à μ ou à σ croissent avec $x_i - \mu$, surtout celle concernant σ ⁽²⁾ de sorte que au sein des Σ les fréquences empiriques extrêmes sont affectées d'un grand poids. Cela n'a rien d'illogique puisque les fréquences éloignées de la moyenne sont très peu probables et qu'elles ont aussi un grand poids dans le calcul de l'écart-type empirique s . Nous aurons l'occasion de revenir sur cet aspect particulier des équations ci-dessus.

2. Pour le moment, et après avoir rappelé les raisons probabilistes du choix de la méthode de détermination des caractéristiques de la population mère, nous exposerons comment on peut, par des considérations d'un autre ordre, aboutir à un choix différent.

Si, abandonnant, momentanément, les raisonnements statistiques, on se tourne vers une conception plus géométrique des choses, on est conduit à énoncer le problème de la manière suivante :

Étant donné une distribution empirique des fréquences d'une variable aléatoire x , déterminer la densité de probabilité de la population mère de manière que les deux répartitions s'écartent dans l'ensemble le moins possible l'une de l'autre. Pour reprendre les notations de l'exposé qui précède, choisir la loi de distribution des probabilités p_i de manière que l'ensemble des écarts $f_i - np_i$ soit aussi faible que possible.

Or on constate que

$$\sum_{i=1}^{i=g} (f_i - np_i) = 0$$

soit :

$$\sum_{i=1}^{i=g} f_i = n$$

égalité toujours vérifiée qui n'apporte aucune solution au problème.

1. On notera que le fait d'affecter à la valeur x_i milieu de la i ème classe les f_i valeurs de la variable étudiées dans l'intervalle $\left(x_i - \frac{\Delta x}{2}, x_i + \frac{\Delta x}{2}\right)$ constitue aussi et déjà une approximation du même ordre.

2. Voir le renvoi ⁽¹⁾ de la p. 246.

On s'adressera donc aux carrés des écarts en écrivant :

$$\sum_{i=1}^{i=q} (f_i - np_i)^2 \text{ minimum}$$

ce qui conduit à l'annulation des dérivées partielles par rapport à μ et σ .

$$\left\{ \begin{array}{l} \sum_{i=1}^{i=q} (f_i - np_i) \frac{\partial p_i}{\partial \mu} = 0 \\ \sum_{i=1}^{i=q} (f_i - np_i) \frac{\partial p_i}{\partial \sigma} = 0 \end{array} \right. \quad (4)$$

Nous notons immédiatement que le système (4) est différent du système (1). Il ne correspond donc pas aux caractéristiques les plus probables, mais à celles qui entraînent dans l'ensemble les fréquences les plus voisines.

Poursuivant le raisonnement de la même manière que précédemment, on aboutira au système (5) qui se substitue au système (2) :

$$\left\{ \begin{array}{l} \sum_{i=1}^{i=q} f_i \frac{\partial \varphi(x_i)}{\partial \mu} = n \sum_{i=1}^{i=q} \varphi(x_i) \frac{\partial \varphi(x_i)}{\partial \mu} \\ \sum_{i=1}^{i=q} f_i \frac{\partial \varphi(x_i)}{\partial \sigma} = n \sum_{i=1}^{i=q} \varphi(x_i) \frac{\partial \varphi(x_i)}{\partial \sigma} \end{array} \right. \quad (5)$$

Si l'on particularise maintenant la fonction $\varphi(x)$ en lui donnant la forme de la loi de Gauss, on aura un système (6) dont il est possible de calculer exactement les seconds membres qui ne contiennent que la fonction $\varphi(x_i)$. Mais auparavant, il convient de remarquer que le fait d'avoir raisonné sur une répartition en q classes des valeurs de la variable x laisse supposer que q est fini, la première de ces classes correspondant à la première valeur non nulle de la fréquence f , de même que la q^{me} correspond à sa dernière valeur non nulle. Le raisonnement qui précède s'est donc limité à l'intervalle correspondant aux q classes. Autrement dit, en raisonnant comme on vient de le faire, on a négligé les écarts des 2 répartitions situés en dehors de l'intervalle (1^{re} classe, q^{me} classe). Il convient de corriger cette erreur qui n'affecte d'ailleurs que le second membre. En effet, dans le premier membre, les facteurs f_i sont nuls en dehors dudit intervalle. Dans le second membre, le rétablissement des termes négligés permet d'ailleurs une simplification : on supposera, en effet, que Δx est suffisamment petit pour que l'on puisse remplacer la somme Σ par l'intégrale correspondante. Les conditions que nous voulons imposer à la recherche de μ et de σ s'écrivent alors,

$$\begin{aligned} \sum_{i=1}^{i=q} f_i \frac{\partial \varphi(x_i)}{\partial \mu} &= n \int_{-\infty}^{+\infty} \varphi(x) \frac{\partial \varphi(x)}{\partial \mu} dx \\ \sum_{i=1}^{i=q} f_i \frac{\partial \varphi(x_i)}{\partial \sigma} &= n \int_{-\infty}^{+\infty} \varphi(x) \frac{\partial \varphi(x)}{\partial \sigma} dx \end{aligned} \quad (5 \text{ bis})$$

Mais il est facile de calculer la valeur des seconds membres : on a en effet :

$$\frac{\partial \varphi}{\partial \mu} = \frac{-1}{\sigma^3 \sqrt{2\pi}} (x - \mu) e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\frac{\partial \varphi}{\partial \sigma} = \frac{1}{\sigma^2 \sqrt{2\pi}} \left[\frac{(x - \mu)^2}{\sigma^2} - 1 \right] e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

La fonction φ étant paire par rapport à l'ordonnée d'abscisse μ et la fonction $\frac{\partial \varphi}{\partial \mu}$ étant impaire, le produit $\varphi \frac{\partial \varphi}{\partial \mu}$ est impair et l'intégrale de $-\infty$ à $+\infty$ est nulle.

La première condition se réduit donc à :

$$\sum_{i=1}^{i=q} f_i \frac{\partial \varphi(x_i)}{\partial \mu} = 0$$

Le second membre de la deuxième équation n'est pas nul. Il se calcule sans difficulté et la seconde équation s'écrit :

$$\sum_{i=1}^{i=q} f_i \frac{\partial \varphi(x_i)}{\partial \sigma} = \frac{-n}{4\sigma^2 \sqrt{\pi}}$$

ou encore

$$\begin{cases} \sum_{i=1}^{i=q} f_i (x_i - \mu) e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} = 0 \\ \sum_{i=1}^{i=q} f_i \left[1 - \frac{(x_i - \mu)^2}{\sigma^2} \right] e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} = \frac{n}{2\sqrt{2}} \end{cases} \quad (6)$$

La résolution du système se fait par approximations successives. A cet effet, on peut partir des valeurs de la moyenne m et de l'écart type s de la distribution étudiée. On trouvera pour les premiers membres des 2 équations des valeurs respectivement différentes de 0 et $\frac{n}{2\sqrt{2}}$. Il conviendra alors de faire un 2^e essai avec des valeurs de μ et σ choisies de manière à changer le signe du premier membre (première équation) et à encadrer la valeur du second membre (2^e équation).

Lorsqu'on aura obtenu ce résultat, une interpolation linéaire donnera en général une approximation suffisante. Très souvent, l'examen des représentations graphiques superposées des fréquences empiriques d'une part, de la courbe de Gauss de caractéristiques m et s d'autre part, pourra servir de guide pour le choix des paramètres du 2^e essai.

A ce qui précède, nous ajouterons la remarque suivante. Les répartitions empiriques se présentent évidemment d'une manière discontinue comme nous l'avons décrit précédemment, mais rien n'empêche de s'intéresser au cas où la répartition étudiée serait susceptible

d'être représentée par une fonction continue $y(x)$: on peut même dire que le raisonnement se simplifie notablement pour aboutir finalement aux 2 équations :

$$\int_{-\infty}^{+\infty} y(x - \mu) e^{-\frac{(x - \mu)^2}{2\sigma^2}} dx = 0$$

$$\int_{-\infty}^{+\infty} y \left[1 - \frac{(x - \mu)^2}{\sigma^2} \right] e^{-\frac{(x - \mu)^2}{2\sigma^2}} dx = \frac{1}{2\sqrt{2}}$$

Si cette remarque n'a pas d'application pratique, elle présente par contre l'intérêt de permettre un test rigoureux de la méthode préconisée. On peut, en effet, choisir une fonction $y(x)$ relativement simple, c'est à dire permettant l'intégration des premiers termes et par conséquent le calcul de μ et de σ en fonction de m et de s .

On peut choisir, à cet effet, la fonction y représentée par les polynômes suivants :

de $-\infty$ à $-a$, $y = 0$
 de $-a$ à $-\frac{a}{2}$, $y = \frac{2}{a} \left(1 + \frac{x}{a} \right)^2$
 de $-\frac{a}{2}$ à $+\frac{a}{2}$, $y = \frac{a^2 - 2x^2}{a^3}$
 de $+\frac{a}{2}$ à $+a$, $y = \frac{2}{a} \left(1 - \frac{x}{a} \right)^2$
 de $+a$ à $+\infty$, $y = 0$

Naturellement, y est telle que son intégrale prise de $-\infty$ à $+\infty$ est égale à l'unité. D'autre part, dans ce cas particulier, comme il y a symétrie par rapport à $x = 0$, la recherche de μ ne se pose pas et l'on a : $\mu = m = 0$. Au terme d'un calcul détaillé en annexe I, on trouve

$$\sigma = 0,384 a$$

que l'on compare à l'écart type de la fonction, $s = 0,354 a$.

La représentation graphique illustre bien les résultats obtenus.

Comme exemple de répartition empirique, on donne en annexe II une distribution de moyenne $m = 100$ et d'écart-type $s = 10,82$.

Si, dans une première approximation, on néglige la recherche de μ , c'est-à-dire si on suppose $\mu = m$, on obtient par interpolation entre deux essais $\sigma = 7,37$ chiffre nettement inférieur à $s = 10,82$. Le report sur un même graphique de la courbe des fréquences empiriques de la courbe de Gauss dont l'écart type est égal à celui de la distribution empirique ($s = 10,82$) et de la courbe de Gauss dont l'écart type est de 7,37, montre combien cette dernière est plus satisfaisante que l'autre.

3. Sur le plan théorique, il paraît y avoir intérêt à comparer les résultats obtenus par les deux méthodes : Pour effectuer cette comparaison, nous aurons recours au test du χ^2 .

Rappelons que le test du χ^2 ou critère de Pearson assigne une certaine borne à la quantité qui a pour expression

$$\sum_{i=1}^{i=q} \frac{(f_i - np_i)^2}{np_i}$$

Si cette quantité est inférieure à la borne en question, la distribution empirique f_i est probablement issue de la répartition théorique p_i avec un risque d'erreur inférieur à celui que l'on s'est fixé. On satisfera donc d'autant plus aisément au critère de Pearson que la

répartition p_i que l'on aura choisie minimisera la somme ci-dessus, ce qui s'exprime au moyen des dérivées partielles par les équations ci dessous (1)

$$\sum_{i=1}^{i=q} \frac{f_i^2}{p_i^2} \frac{\partial p_i}{\partial \mu} = 0$$

$$\sum_{i=1}^{i=q} \frac{f_i^2}{p_i^2} \frac{\partial p_i}{\partial \sigma} = 0$$
(7)

On se trouve ainsi pour déterminer μ et σ en présence d'un nouveau système de 2 équations que l'on peut comparer aux systèmes (1) et (4) exprimant respectivement le maximum de vraisemblance et le minimum de la somme des carrés des écarts.

La confrontation des 3 systèmes est relativement facile. En effet, d'une part, on obtient les termes du système (7) en multipliant ceux du système (1) par $\frac{f_i}{p_i}$, d'autre part, ceux du système (4) sont ceux du système (1) multipliés par p_i .

En un mot, c'est le système (7) qui donne le plus grand poids aux fréquences extrêmes. Vient ensuite le système (4). Pour fixer les ordres de grandeur, choisissons en effet, pour p_i une répartition de Gauss. Les termes du système (1) sont de la forme [voir (3)] :

$$f_i (x_i - \mu)$$

$$f_i \left[-\frac{1}{\sigma} + \frac{(x_i - \mu)^2}{\sigma^3} \right]$$

tandis que ceux du système (4) s'écrivent [voir (6)] :

$$f_i (x_i - \mu) e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$

$$f_i \left[1 - \frac{(x_i - \mu)^2}{\sigma^2} \right] e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$

Lorsque $x_i - \mu$ est grand par rapport à σ (termes extrêmes), ceux du système (4) sont très petits et deviennent même négligeables devant ceux du système (1). Quant au système (7), il multiplie par $\frac{f_i}{p_i}$ le poids des termes du système (1) : autant dire que le rapport (7)/(1) est équivalent au rapport (1)/(4) en ce qui concerne l'ordre de grandeur des termes extrêmes.

Il est ainsi possible d'affirmer qu'en adoptant le système (4) qui correspond à la recherche du minimum de la somme des carrés des écarts, on atténue le rôle joué par les fréquences extrêmes de la répartition empirique. En particulier, si comme dans l'exemple précédent, il existe des fréquences isolées très éloignées du centre, l'emploi du système (4) équivaut non plus à en atténuer le rôle, mais même à les négliger. C'est finalement une manière d'éliminer les fréquences anormales.

1. La dérivation pure et simple donne $\sum \frac{f_i^2 - n^2 p_i^2}{p_i^2} \frac{\partial p_i}{\partial \mu} = 0$ mais cette équation se ramène à celle du texte si l'on n'oublie pas que $\sum p_i$ étant égal à n , $\sum \frac{\partial p_i}{\partial \mu}$ est égal à zéro.

On remarquera que, quoiqu'il en soit, le résultat obtenu par le système (4) ne peut jamais être aberrant : en effet, les 3 systèmes sont rigoureusement identiques lorsqu'il existe une répartition théorique $\varphi(x)$ telle que pour toute valeur x_i de l'échantillon on ait $p_i = f_i$. Cette constatation ne manque pas d'être rassurante.

4. Reprenant le système d'équation de la page 246 qui définissait le choix de la fonction $\varphi(x)$ par la recherche du maximum de vraisemblance, et constatant que lorsque l'on donne à la fonction $\varphi(x)$ la forme de la loi de Gauss, on aboutit à l'identification des moyennes et des écarts-types, on est amené à se demander si le système en question conduit inmanquablement à ce résultat quelle que soit la forme donnée à la fonction.

La réponse est négative. On ne trouve rien, d'ailleurs, dans le système en question qui permette de le supposer. Mais on peut tenter de définir la forme à donner à la fonction $\varphi(x)$ pour obtenir le résultat en question.

Soit donc $\varphi(x, \mu, \sigma)$ la fonction cherchée. Si le système de la page 246 conduit à l'identification des moyennes et des écarts-types, c'est qu'il équivaut à :

$$\left\{ \begin{array}{l} \mu = \sum \frac{x_i}{n} \\ \sigma^2 = \sum \frac{(x_i - \mu)^2}{n} \end{array} \right.$$

Appelons g_1, g_2, h_1, h_2 , quatre fonctions de μ et de σ : les deux systèmes seront équivalents si l'on peut écrire

$$\left\{ \begin{array}{l} \sum \frac{\partial}{\partial \mu} \log \varphi(x) = g_1 \Sigma (x_i - \mu) + g_2 \Sigma [(x_i - \mu)^2 - \sigma^2] \\ \sum \frac{\partial}{\partial \sigma} \log \varphi(x) = h_1 \Sigma (x_i - \mu) + h_2 \Sigma [(x_i - \mu)^2 - \sigma^2] \end{array} \right.$$

et comme ces équations doivent être vérifiées quel que soit x , les deux systèmes seront équivalents si l'on peut écrire

$$\frac{\partial}{\partial \mu} \log \varphi = (x - \mu) g_1 + [(x - \mu)^2 - \sigma^2] g_2$$

et l'égalité qui s'en déduit en dérivant par rapport à σ et en remplaçant g par h .

Mais ces deux équations ne seront compatibles que s'il y a identité des dérivées secondes

$\frac{\partial^2}{\partial \mu \partial \sigma} \log \varphi$ et cela aussi quel que soit x

$$(x - \mu) \frac{\partial g_1}{\partial \sigma} + [(x - \mu)^2 - \sigma^2] \frac{\partial g_2}{\partial \sigma} - 2 \sigma g_2 \equiv - h_1 + (x - \mu) \frac{\partial h_1}{\partial \mu} + [(x - \mu)^2 - \sigma^2] \frac{\partial h_2}{\partial \mu} - 2 (x - \mu) h_2$$

Écrivons que les coefficients des termes en $x - \mu$, $(x - \mu)^2$, et que les termes constants sont égaux dans les 2 membres; on obtient les 3 relations suivantes entre les 4 fonctions de départ :

$$\left\{ \begin{array}{l} h_1 = 2 \sigma g_2 \\ \frac{\partial g_1}{\partial \sigma} = \frac{\partial h_1}{\partial \mu} - 2 h_2 \\ \frac{\partial g_2}{\partial \sigma} = \frac{\partial h_2}{\partial \mu} \end{array} \right.$$

Il paraît tout indiqué de prendre g_2 comme fonction arbitraire. g_1 , h_1 , h_2 s'écriront alors :

$$\left\{ \begin{array}{l} g_1 = 2 \left[\int \sigma \frac{\partial g_2}{\partial \mu} d\sigma - \int g_2 d\mu \right] \\ h_1 = 2 \sigma g_2 \\ h_2 = \int \frac{\partial g_2}{\partial \sigma} d\mu \end{array} \right.$$

Ceci permet de récrire les équations de départ, d'où l'on tirera la fonction $\varphi(x)$ sous la forme

$$\left\{ \begin{array}{l} \frac{\partial}{\partial \mu} \log \varphi = 2(x - \mu) \left[\int \sigma \frac{\partial g_2}{\partial \mu} d\sigma - \int g_2 d\mu \right] + [(x - \mu)^2 - \sigma^2] g_2 \\ \frac{\partial}{\partial \sigma} \log \varphi = 2 \sigma g_2 (x - \mu) + [(x - \mu)^2 - \sigma^2] \int \frac{\partial g_2}{\partial \sigma} d\mu \end{array} \right.$$

Il faut avoir soin de noter que les intégrales ne sont définies qu'à une constante près par rapport à la variable d'intégration, c'est à dire qu'à une fonction arbitraire près de l'autre variable. Un cas particulier intéressant illustrera cette remarque. Examinons l'hypothèse $g_2 = 0$. On a évidemment aussi

$$\frac{\partial g_2}{\partial \mu} = \frac{\partial g_2}{\partial \sigma} = 0$$

mais
$$\int \sigma \frac{\partial g_2}{\partial \mu} d\sigma, \quad \int g_2 d\mu \quad \text{et} \quad \int \frac{\partial g_2}{\partial \sigma} d\mu$$

ne sont pas nulles.

Appelons $\psi_1(\mu)$, $\psi_2(\sigma)$, $\psi_3(\sigma)$, les fonctions en question : le cas étudié ($g_2 = 0$) conduit à

$$\left\{ \begin{array}{l} \frac{\partial}{\partial \mu} \log \varphi = 2(x - \mu) [\psi_1(\mu) + \psi_2(\sigma)] \\ \frac{\partial}{\partial \sigma} \log \varphi = [(x - \mu)^2 - \sigma^2] \psi_3(\sigma) \end{array} \right.$$

On constate d'ailleurs immédiatement que l'identité des dérivées secondes $\frac{\partial^2}{\partial \mu \partial \sigma} \log \varphi(x)$

exige que $\frac{d\psi_2}{d\sigma} = -\psi_3$

Le cas particulier $g_2 = 0$ aboutit donc à

1^{re} équation sans changement

$$\left\{ \frac{\partial}{\partial \sigma} \log \varphi = -[(x - \mu)^2 - \sigma^2] \frac{d\psi_2}{d\sigma} \right.$$

La même remarque appliquée à l'intégration de ces deux équations montre rapidement que $\psi_1(\mu)$ est nécessairement identiquement nulle, si bien que la fonction $\log \varphi$ est de la forme

$$- (x - \mu)^2 \psi(\sigma) + \int \sigma^2 \frac{d\psi}{d\sigma} d\sigma + C''$$

Si l'on choisit $\psi(\sigma) = \frac{1}{2\sigma^2}$ on retrouve la loi de Gauss.

Ainsi, nous avons montré quelles formes devait revêtir la fonction $\varphi(x)$ pour que la recherche du maximum de vraisemblance équivale à l'identification des moyennes et des écarts types. Il est possible de donner un exemple de fonction pour laquelle la recherche du maximum de vraisemblance n'aboutit pas à ce résultat.

Prenons en effet la loi de densité de probabilité représentée par la fonction

$$y = \frac{2\sigma^3}{\pi} \left[\frac{1}{\sigma^2 + (x_i - \mu)^2} \right]^2$$

On connaît la primitive de la fonction ⁽¹⁾ $\frac{1}{(1+t^2)^2}$ qui est $\frac{1}{2} \left(\text{Arc tg } t + \frac{t}{1+t^2} \right)$

il est donc aisé de vérifier que $\int_{-\infty}^{+\infty} \varphi(x) dx$ est bien égal à l'unité.

La moyenne par raison de symétrie est égale à μ et l'écart type est égal à σ .

Soit une répartition empirique consistant en un échantillon de 4 valeurs de la variable :

$$x_1 = 1 \quad x_2 = 3 \quad x_3 = 5 \quad x_4 = 6$$

La moyenne est égale à 3,75 et l'écart type est égal à 1,92...

Écrivons d'autre part que la population-mère répond au système d'équations (2) dans lesquelles on a

$$f_1 = f_2 = f_3 = f_4 = 1$$

Ces équations deviennent :

$$\sum \frac{x_i - \mu}{\sigma^2 + (x_i - \mu)^2} = u(\mu, \sigma) = 0$$

$$\sum \frac{\sigma^2}{\sigma^2 + (x_i - \mu)^2} - 3 = v(\mu, \sigma) = 0$$

Pour résoudre ce système, on trace sur un même graphique les courbes représentant

$$u(\mu, \sigma) = 0 \quad \text{et} \quad v(\mu, \sigma) = 0$$

leur intersection donne la solution cherchée :

$$\mu = 3,93$$

$$\sigma = 3,025$$

On voit que l'on est assez loin de la moyenne et de l'écart-type de l'échantillon.

1. On sait qu'il n'est pas possible d'utiliser la fonction $\frac{1}{1+t^2}$ qui n'a ni moyenne ni écart type.

5. Après nous être beaucoup référé à la loi de Gauss, il nous paraît indispensable de compléter ce qui précède par quelques considérations relatives à la recherche d'une forme de densité de probabilité essentiellement dissymétrique.

La plupart des traités élémentaires de calcul des probabilités donnent comme lois fondamentales de distribution des probabilités d'une variable aléatoire la loi binomiale, les lois de Poisson et de Gauss. Ces trois lois peuvent donner lieu aux remarques suivantes :

- l'une d'entre elles seulement est continue et donne la densité de probabilité de la variable qui peut prendre (au moins en principe) n'importe quelle valeur de $-\infty$ à $+\infty$; les autres sont discontinues donnant la probabilité de réalisation des valeurs entières de la variable comprises dans un intervalle positif donné (cet intervalle s'étendant jusqu'à $+\infty$ pour la loi de Poisson).
- l'une d'entre elles seulement possède un axe de symétrie tandis que les deux autres sont en général dissymétriques (la distribution binomiale n'est symétrique que pour la valeur 0,5 du paramètre).

Des deux remarques qui précèdent, on conclut que la théorie élémentaire ne présente pas de fonction de distribution continue dissymétrique. Or, on est bien obligé de constater que nombre de variables aléatoires peuvent varier d'une manière continue en ne couvrant par essence même qu'une partie du domaine possible de variation. C'est le cas par exemple de l'âge d'un individu, d'une hauteur de précipitation atmosphérique ou du débit d'un cours d'eau qui ne sauraient ni les uns ni les autres, prendre des valeurs négatives, alors que leurs variations peuvent être parfaitement continues.

Il semble que ce soit cette notion de continuité qui a donné à la loi de Gauss l'importance que l'on connaît, faisant oublier — ou tout au moins négliger — le fait qu'elle est symétrique et s'étend de $-\infty$ à $+\infty$ alors que les variables aléatoires auxquelles on veut l'appliquer ne peuvent donner lieu très souvent qu'à des distributions dissymétriques correspondant aux seules valeurs positives.

On paraît aussi perdre de vue que la loi de Gauss est initialement une loi de répartition des erreurs. On a même fait la démonstration que lorsque l'on a affaire à un grand nombre de mesures, donc à un grand nombre d'erreurs qui se compensent, ces erreurs suivent la loi de Gauss. Nous ne rappellerons pas cette démonstration, mais nous croyons qu'il était utile de souligner que la loi de Gauss est essentiellement adaptée à des répartitions de probabilité d'erreurs, c'est-à-dire, notamment, de variables dont l'ordre de grandeur reste petit par rapport à la grandeur mesurée, l'ajustement n'étant satisfaisant que lorsque le nombre n au contraire est grand.

En fait de loi dissymétrique élémentaire, on a donc à disposition la loi de Poisson et la loi binomiale lesquelles ne s'appliquent directement qu'à des variables aléatoires discontinues.

Pour disposer de lois de répartition dissymétriques applicables à des variables aléatoires continues, il faut considérer les lois plus complexes γ , β , log normale, χ^2 , t de Student ou F de Fischer-Snedecor.

6. Dans ce domaine, nous signalons une possibilité qui ne paraît pas avoir été exploitée et qui consiste à utiliser une interpolation de la loi de Poisson.

Au lieu de son expression traditionnelle :

$$f(k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

qui n'est d'ailleurs pas nulle pour $k = 0$, nous écrivons :

$$f(x) = e^{-\lambda} \frac{\lambda^{x-1}}{\Gamma(x)}$$

la fonction $\Gamma(x)$ étant celle qui interpole $(x - 1)!$. Cette fois, x peut prendre toutes les valeurs positives ou nulles et la valeur de $f(x)$ pour $x = 0$ est nulle.

Maintenant, on dispose d'une loi de répartition qui non seulement peut s'appliquer à une variable aléatoire continue, mais qui dépend de deux paramètres (a et λ). Elle peut en effet s'écrire :

$$f(x) = \frac{e^{-\lambda}}{a} \frac{\lambda^{\frac{x}{a}-1}}{\Gamma\left(\frac{x}{a}\right)}$$

Il faut tout de suite noter que pour que $f(x)$ représente bien une loi de densité de probabilité, il faut que l'intégrale :

$$\int_0^{\infty} f(x) dx \text{ soit égale à l'unité}$$

Or, rien ne permet de supposer que ce soit bien le cas. La somme des ordonnées de la fonction en question prises d'unité en unité est bien égale à 1 et par conséquent, la ligne polygonale qui joint les sommets des trapèzes ainsi formés enferme bien une surface unité entre elle et l'axe des x . Mais il n'en est pas ainsi de la courbe représentative de la fonction continue.

Il y a donc lieu de calculer l'intégrale $I(\lambda)$ définie par

$$I(\lambda) = \int_0^{\infty} e^{-\lambda} \frac{\lambda^{x-1}}{\Gamma(x)} dx$$

Appelons $y(x)$ la fonction à intégrer : cette fonction est telle que la méthode des trapèzes appliquée avec des abscisses équidistantes d'une unité donne une première approximation égale à l'unité; on a, en effet,

$$\sum_{i=0}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} = e^{\lambda}$$

On sait qu'une telle approximation peut ensuite être corrigée par les termes suivants du développement d'Euler-MacLaurin qui fait intervenir la variation d'une extrémité à l'autre de l'intervalle d'intégration des dérivées impaires successives de la fonction. Comme toutes ces dérivées sont nulles quand $x \rightarrow \infty$, les variations correspondantes sont égales aux valeurs prises pour $x = 0$ des dérivées changées de signe. Et l'on a :

$$I = 1 + \frac{1}{12} y_0' - \frac{1}{720} y_0''' + \dots - \frac{(-1)^n}{(2n)!} B_n y_0^{(2n-1)} + \dots$$

où $B_1 B_2 \dots B_n$ sont les nombres de Bernoulli ⁽¹⁾.

Il suffit donc de chercher y'_0, y''_0, \dots etc. La méthode que nous avons utilisée repose sur l'identification du développement en série de Taylor de la fonction y , qui s'écrit évidemment

$$y = y'_0 \cdot x + y''_0 \frac{x^2}{2!} + y'''_0 \frac{x^3}{3!} + \dots$$

et des développements de chacun de ses deux facteurs pris séparément

$$\lambda^{x-1} = \frac{1}{\lambda} \left(1 + \frac{x \log \lambda}{1!} + \frac{x^2 \log^2 \lambda}{2!} + \dots \right) \quad \text{développement connu}$$

$$\frac{1}{\Gamma(x)} = x + \alpha x^2 + \beta x^3 + \dots \quad \text{à déterminer}$$

Le détail des calculs est donné en annexe III. Il aboutit aux résultats suivants :

$$\lambda e^\lambda y'_0 = 1$$

$$\lambda e^\lambda y''_0 = 2 (\log \lambda - \varphi_1)$$

$$\lambda e^\lambda y'''_0 = 3 [(\log \lambda - \varphi_1)^2 - \varphi'_1]$$

$$\lambda e^\lambda y^{IV}_0 = 4 [(\log \lambda - \varphi_1)^3 - 3(\log \lambda - \varphi_1) \varphi'_1 - \varphi''_1]$$

$$\lambda e^\lambda y^V_0 = 5 [(\log \lambda - \varphi_1)^4 - 6(\log \lambda - \varphi_1)^2 \varphi'_1 - 4(\log \lambda - \varphi_1) \varphi''_1 + 3\varphi'^2_1 - \varphi'''_1]$$

Les paramètres $\varphi_1, \varphi'_1, \dots$ sont les valeurs prises pour $x = 1$ par la dérivée logarithmique de $\Gamma(x)$ et ses dérivées successives qui ont les valeurs suivantes :

$$\varphi_1 = -0,5772157$$

$$\varphi'_1 = +1,6449341$$

$$\varphi''_1 = -2,4041138$$

$$\varphi'''_1 = +6,4939394$$

On remarquera d'autre part que, selon les propriétés de la série d'Euler-Mac Laurin, la convergence n'est pas assurée jusqu'à l'infini; mais on sait que cette série commence toujours par converger.

On obtient ainsi par exemple :

$$I \left(\frac{1}{2} \right) = 1,1094$$

$$I (1) = 1,0329$$

$$I (2) = 1,0056$$

$$I (3) = 1,0013$$

⁽¹⁾ $B_1 = \frac{1}{6}, B_2 = \frac{1}{30}, B_3 = \frac{1}{42}, B_4 = \frac{1}{30}, B_5 = \frac{5}{66} \dots$

Nous ajouterons les considérations suivantes sur le calcul de la moyenne et des moments des divers types. Appelons M la moyenne, S le moment du second ordre, auxquels nous nous limiterons ici. On a :

$$I(\lambda) \cdot M(\lambda) = e^{-\lambda} \int_0^{\infty} \frac{x \lambda^{x-1}}{\Gamma(x)} dx$$

$$I(\lambda) \cdot S(\lambda) = e^{-\lambda} \int_0^{\infty} \frac{x^2 \lambda^{x-1}}{\Gamma(x)} dx$$

D'une part l'évaluation par la méthode des trapèzes ayant l'unité pour base donne pour premières approximations des seconds membres respectivement :

$$\begin{aligned} 1 + \lambda \\ 1 + 3\lambda + \lambda^2 \end{aligned}$$

D'autre part, la même identification que précédemment donne pour valeurs des dérivées :

pour la moyenne

$$\begin{aligned} y_0 &= y'_0 = 0 \\ y''_0 &= 2 \lambda^{-1} e^{-\lambda} \\ y'''_0 &= 6 \lambda^{-1} e^{-\lambda} (\log \lambda - \varphi_1) \\ y^{IV}_0 &= 12 \lambda^{-1} e^{-\lambda} [(\log \lambda - \varphi_1)^2 - \varphi'_1] \\ y^V_0 &= 20 [(\log \lambda - \varphi_1)^3 - 3(\log \lambda - \varphi_1) \varphi'_1 - \varphi''_1] \end{aligned}$$

pour le moment du second ordre :

$$\begin{aligned} y_0 &= y'_0 = y''_0 = 0 \\ y'''_0 &= 6 \lambda^{-1} e^{-\lambda} \\ y^{IV}_0 &= 24 \lambda^{-1} e^{-\lambda} (\log \lambda - \varphi_1) \\ y^V_0 &= 60 \lambda^{-1} e^{-\lambda} [(\log \lambda - \varphi_1)^2 - \varphi'_1] \\ y^{VI}_0 &= 210 \lambda^{-1} e^{-\lambda} [(\log \lambda - \varphi_1)^3 - 6(\log \lambda - \varphi_1)^2 \varphi'_1 - 4(\log \lambda - \varphi_1) \varphi''_1 + 3 \varphi'_1 - \varphi'''_1] \end{aligned}$$

Dans les 2 cas, les calculs des dérivées successives sont exactement les mêmes à un facteur constant près. C'est ainsi que l'on trouve :

$$\begin{array}{ll} M\left(\frac{1}{2}\right) = 1,3549 & S\left(\frac{1}{2}\right) = 2,4640 \\ M(1) = 1,9345 & S(1) = 4,8368 \\ M(2) = 2,9824 & S(2) = 10,9374 \\ M(3) = 3,9946 & S(3) = 18,9752 \end{array}$$

et les variances s'en déduisent par $V = S - M^2$ puis les écarts-types par $\sigma = \sqrt{V}$

$$\begin{array}{ll} V\left(\frac{1}{2}\right) = 0,6282 & \sigma\left(\frac{1}{2}\right) = 0,7926 \\ V(1) = 1,0945 & \sigma(1) = 1,04619 \\ V(2) = 2,0427 & \sigma(2) = 1,4295 \\ V(3) = 3,0186 & \sigma(3) = 1,7374 \end{array}$$

Enfin, on peut calculer le mode encore plus facilement : c'est la valeur de la variable qui annule la dérivée de la fonction $\frac{\lambda^{x-1}}{\Gamma(x)}$. Or, cette dérivée est égale à : $\frac{\lambda^{x-1}}{\Gamma(x)} \left[\log \lambda - \frac{\Gamma'(x)}{\Gamma(x)} \right]$

La recherche du mode conduit donc à la recherche des solutions de l'équation :

$$\varphi(x) = \log \lambda$$

Les considérations développées dans l'annexe montrent qu'il est très facile de calculer avec beaucoup de précision les valeurs de $\varphi(x)$ pour n'importe quelle valeur de x . La valeur cherchée s'obtient ensuite par interpolation. La même annexe donne le tableau des valeurs de $\varphi(x)$ pour x variant de 1 à 2 de vingtième en vingtième. Voici les valeurs qui s'en déduisent pour l'abscisse et la valeur du mode :

λ	x_M	y_M	$y_M/I(\lambda)$
0,5	0,9335...	0,6089...	0,5489...
1	1,4618...	0,4154...	0,4022...
2	2,4798...	0,2880...	0,2864...
3	3,4863...	0,2335...	0,2332...

L'intérêt de ce dernier calcul provient du fait qu'il rend possible un tracé précis des courbes correspondant aux différentes valeurs de λ . Pour que la comparaison avec une courbe empirique soit possible et pour que l'on détermine avec facilité la valeur de λ qui correspond le mieux à la répartition étudiée, nous suggérons de multiplier les abscisses de chaque courbe par un coefficient tel que leurs modes viennent en coïncidence. A ce moment, les ordonnées correspondantes devront être divisées par le même coefficient pour que la surface comprise entre l'axe des abscisses et chacune d'entre elles reste égale à l'unité. Si l'on appelle u la nouvelle variable commune prenant la valeur u_M lorsque $x = x_M$, elle doit être telle que

$$\frac{u}{u_M} = \frac{x}{x_M}$$

Dans la fonction précédente, il faut donc remplacer x par $\frac{ux_M}{u_M}$. L'ordonnée subira la transformation inverse et deviendra z telle que $y = \frac{zu_M}{x_M}$ d'où :

$$z = \frac{x_M}{u_M} \frac{e^{-\lambda}}{I_\lambda} \frac{\lambda^{\frac{ux_M}{u_M} - 1}}{\Gamma\left(\frac{ux_M}{u_M}\right)}$$

u_M peut être choisi arbitrairement et x_M fonction de λ prend les valeurs données précédemment.

On constate d'ailleurs que par cette transformation, l'étalement des courbes est d'autant plus grand que λ est plus petit, le maximum z_M qui correspond à u_M variant dans le même sens que λ .

En possession du réseau ci-dessus, il devient facile d'y superposer la courbe représentative de la densité de probabilité étudiée. Il faut pour cela, et en premier lieu, donner à cette dernière, par un changement d'échelle approprié, le même mode que celui du réseau. On en déduit facilement la valeur y_M correspondante ainsi que toutes les valeurs de y .

Une simple superposition permet de se rendre compte si le résultat obtenu est satisfaisant. Dans la négative, on peut assez facilement faire la retouche nécessaire à l'obtention d'une bonne insertion qui permettra par interpolation le calcul des éléments de la courbe théorique.

On voit que cette méthode d'insertion visuelle repose au départ sur l'identification par les modes.

Tel nous paraît être l'intérêt des recherches effectuées sur la fonction proposée : en présence d'une variable aléatoire susceptible de prendre n'importe quelle valeur positive ou nulle, la recherche d'une loi théorique de distribution des densités de probabilité peut s'obtenir au moyen d'une adaptation convenable de la loi de Poisson.

ANNEXE I

L'objet de cette annexe est de déterminer la courbe de Gauss la plus voisine de la loi de densité de probabilité définie par les 3 arcs de paraboles

$$\text{de } -a \text{ à } -\frac{a}{2} \quad y = \frac{2}{a} \left(1 + \frac{x}{a}\right)^2$$

$$\text{de } -\frac{a}{2} \text{ à } +\frac{a}{2} \quad y = \frac{a^2 - 2x^2}{a^3}$$

$$\text{de } +\frac{a}{2} \text{ à } +a \quad y = \frac{2}{a} \left(1 - \frac{x}{a}\right)^2$$

En dehors de l'intervalle $(-a, +a)$ la fonction $y(x)$ est supposée nulle. On vérifie facilement que

$$\int_{-\infty}^{+\infty} y(x) dx = 1$$

On cherche l'écart type σ de la courbe de Gauss solution de l'équation

$$\int_{-\infty}^{+\infty} y \left(1 - \frac{x^2}{\sigma^2}\right) e^{-\frac{x^2}{2\sigma^2}} dx = \frac{1}{2\sqrt{2}}$$

C'est en effet des 2 équations du système de la p. 251, la seule qui reste, la première étant automatiquement satisfaite pour raison de symétrie. Pour la même raison de symétrie, on peut aussi l'écrire

$$\int_0^a y \left(1 - \frac{x^2}{\sigma^2}\right) e^{-\frac{x^2}{2\sigma^2}} dx = \frac{1}{4\sqrt{2}}$$

ou encore en remplaçant y par ses valeurs successives :

$$\int_0^{\frac{a}{2}} \frac{a^2 - 2x^2}{a^3} \left(1 - \frac{x^2}{\sigma^2}\right) e^{-\frac{x^2}{2\sigma^2}} dx + \int_{\frac{a}{2}}^a \frac{2}{a} \left(1 - \frac{x}{a}\right)^2 \left(1 - \frac{x^2}{\sigma^2}\right) e^{-\frac{x^2}{2\sigma^2}} dx = \frac{1}{4\sqrt{2}}$$

Ceci conduit à intégrer des fonctions de la forme $x^j e^{-\frac{x^2}{2\sigma^2}}$ que l'on obtient par récurrence. On a en effet

$$I_j = \int x^j e^{-\frac{x^2}{2\sigma^2}} dx = -\sigma^2 x^{j-1} e^{-\frac{x^2}{2\sigma^2}} + (j-1) \sigma^2 I_{j-2}$$

Si l'on appelle I'_j et I''_j l'intégrale I_j prise respectivement entre 0 et $\frac{a}{2}$ et entre $\frac{a}{2}$ et a , on aboutit au terme d'un calcul assez laborieux à la relation

$$16 \sqrt{2} \sigma^2 (I'_0 - I''_0) = a^3$$

avec
$$I'_0 = \int_0^{\frac{a}{2}} e^{-\frac{x^2}{2\sigma^2}} dx \quad \text{et} \quad I''_0 = \int_{\frac{a}{2}}^a e^{-\frac{x^2}{2\sigma^2}} dx$$

et si l'on pose $\frac{x}{\sigma} = \xi$

$$I'_0 = \sigma \int_0^{\frac{a}{2\sigma}} e^{-\frac{\xi^2}{2}} d\xi \quad I''_0 = \sigma \int_{\frac{a}{2\sigma}}^{\frac{a}{\sigma}} e^{-\frac{\xi^2}{2}} d\xi$$

Les intégrales nécessaires sont données par les tables de la loi normale réduite

$$\Pi(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\xi} e^{-\frac{\xi^2}{2}} d\xi$$

si bien que l'on a à résoudre au prix de quelques tâtonnements l'équation

$$\frac{a^3}{\sigma^3} = 32 \sqrt{\pi} [2 \Pi\left(\frac{a}{2\sigma}\right) - \Pi\left(\frac{a}{\sigma}\right) - \Pi(0)]$$

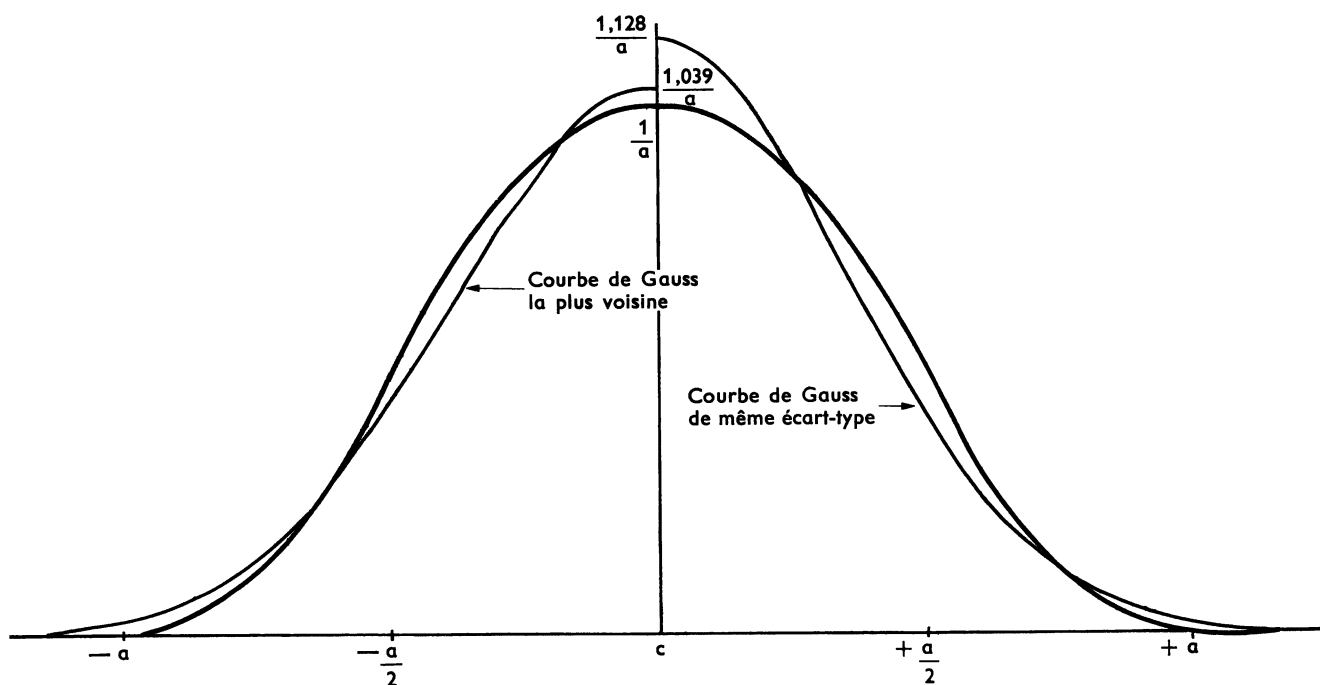
On trouve $\frac{a}{\sigma} = 2,606$ ou encore $\frac{\sigma}{a} = 0,384$

L'écart type de la répartition parabolique d'où l'on est parti est $\frac{a}{2\sqrt{2}} = 0,354 a$.

La courbe correspondant à $\sigma = 0,384 a$ est plus plate que celle qui correspond à l'identification des écarts types. Les sommets qui correspondent à $x = 0$ s'ordonnent de la manière suivante :

courbe étudiée	$\frac{1}{a}$
courbe la plus voisine	$\frac{1,039}{a}$
courbe du même écart-type	$\frac{1,128}{a}$

Ces constatations sont illustrées par la figure 1 où l'on a dessiné les 3 courbes.



— Fig. 1 —

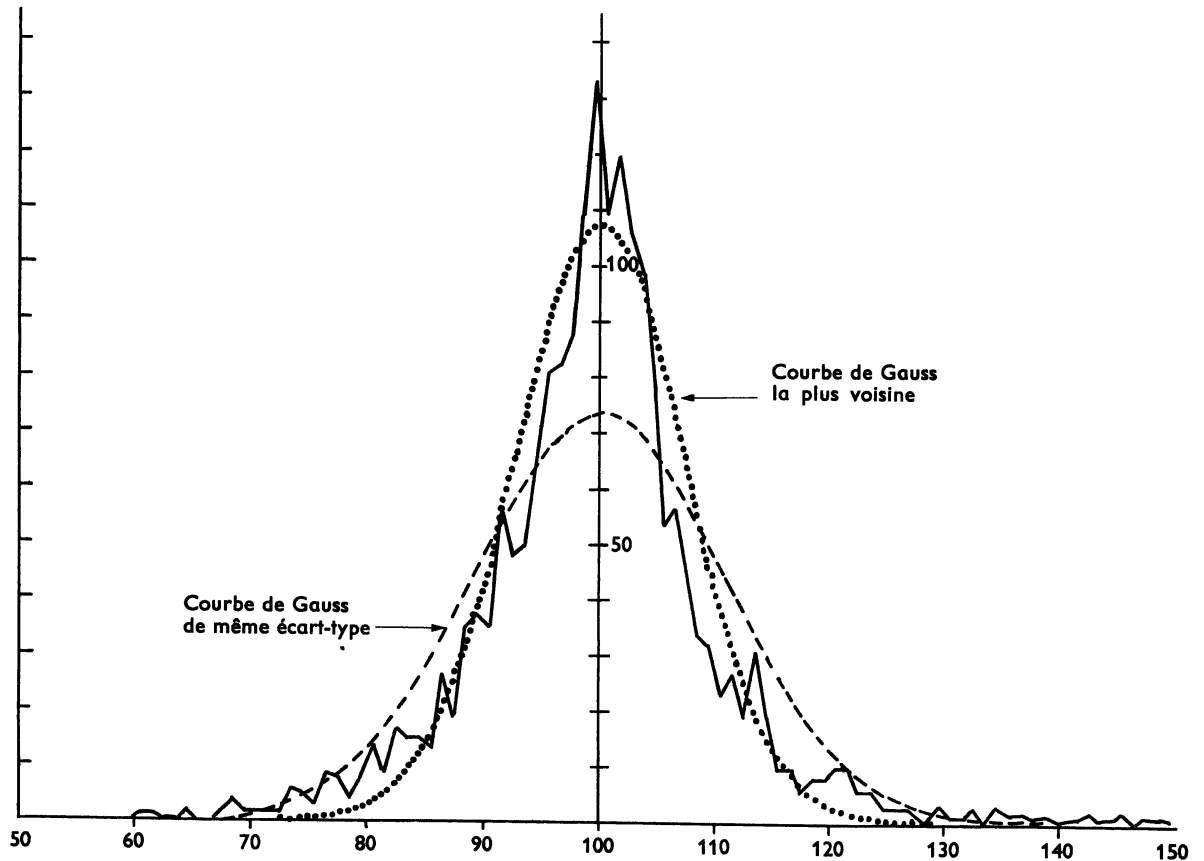
ANNEXE II

La distribution donnée en exemple est la suivante :

Classe	Fréquence	Classe	Fréquence	Classe	Fréquence
45-46	1	93-94	50	128-129	0
.....	0	94-95	67	129-130	3
60-61	1	95-96	81	130-131	1
61-62	1	96-97	83	131-132	1
62-63	0	97-98	87	132-133	3
63-64	0	98-99	109	133-134	0
64-65	2	99-100	133	134-135	3
65-66	0	100-101	109	135-136	2
66-67	0	101-102	120	136-137	2
67-68	2	102-103	106	137-138	1
68-69	4	103-104	99	138-139	1
69-70	2	104-105	77	139-140	0
70-71	2	105-106	54	140-141	1
71-72	2	106-107	57	141-142	1
72-73	2	107-108	44	142-143	2
73-74	6	108-109	34	143-144	1
74-75	5	109-110	32	144-145	1

Classe	Fréquence	Classe	Fréquence	Classe	Fréquence
75-76	3	110-111	23	145-146	2
76-77	9	111-112	27	146-147	0
77-78	8	112-113	19	147-148	1
78-79	4	113-114	31	148-149	1
79-80	8	114-115	19	149-150	1
80-81	14	115-116	10	150-151	0
81-82	9	116-117	10	151-152	1
82-83	17	117-118	6	152-153	0
83-84	15	118-119	8	153-154	1
84-85	15	119-120	8	154-155	1
85-86	13	120-121	10	155-156	0
86-87	27	121-122	10	156-157	0
87-88	19	122-123	6	157-158	0
88-89	35	123-124	6	158-159	1
89-90	37	124-125	3	159-160	0
90-91	35	125-126	3	160-161	1
91-92	58	126-127	3	0
92-93	48	127-128	2	175-176	1

On vérifie que la moyenne $m = 100$ et que l'écart-type $\sigma = 10,82$. Ainsi qu'il est indiqué dans le texte, la distribution paraît grossièrement symétrique par rapport à sa moyenne; on s'est donc contenté de chercher la valeur de σ qui satisfait à la deuxième équation du système (6) p. 250 dans laquelle on porte $\mu = 100$.



— Fig. 2 —

Pour faciliter les calculs, on tient compte du fait que $x_i - \mu$ n'intervient que par son carré; dans le calcul des termes successifs de la somme Σ , on peut donc bloquer les fréquences correspondant à des valeurs opposées de $x_i - \mu$. Par exemple, on calculera le terme correspondant à $x_i - \mu = -25,5$ dont la fréquence est 5 en même temps que celui qui correspond à $x_i - \mu = +25,5$ dont la fréquence est 3. D'autre part, on a intérêt à grouper les termes de même signe et, à ce sujet, l'on remarque que le facteur $1 - \frac{x_i^2}{\sigma^2}$ est positif à l'intérieur de l'intervalle $(-\sigma, +\sigma)$ et négatif à l'extérieur.

Voici les résultats des essais effectués avec $\sigma = 7,2$ et $\sigma = 7,9$. Dans le premier cas, la valeur du premier membre divisé par $n = 1\ 984$ est de 0,3448 et dans le second cas de 0,3808

Ces résultats encadrent la valeur cherchée de $\frac{1}{2\sqrt{2}}$ qui est égale à 0,35355. L'interpolation donne la valeur de σ , égale à 7,37. Sur la figure 2, on a dessiné la courbe de distribution des fréquences, la courbe de Gauss de même écart type ($\sigma = 10,82$) et la courbe de Gauss construite avec l'écart-type de 7,37 issu du calcul qui précède. Cette figure met bien en évidence les différences entre ces deux dernières et permet de les rapprocher l'une et l'autre de la distribution empirique analysée.

ANNEXE III

Il s'agit de calculer l'intégrale

$$I(\lambda) = e^{-\lambda} \int_0^{\infty} \frac{\lambda^{x-1}}{\Gamma(x)} dx$$

et nous avons indiqué que la méthode des trapèzes appliquée à des abscisses croissant de zéro à l'infini d'unité en unité conduisait à une première approximation égale à l'unité. Il suffit alors de corriger cette valeur approchée au moyen des termes suivants de la série d'Euler-Mac Laurin qui mettent en jeu, comme on le sait, les dérivées successives d'ordre impair de la fonction à intégrer dont il faut calculer les valeurs aux deux extrémités de l'intervalle d'intégration. Cet intervalle s'étend de zéro à l'infini; mais on peut facilement constater que toutes ces dérivées sont nulles à l'infini et il suffit d'obtenir leurs valeurs à l'origine. Or ce sont celles qui figurent dans les termes successifs du développement de la fonction en série de Taylor au voisinage de zéro

$$y = y_0 + y'_0 \frac{x}{1!} + y''_0 \frac{x^2}{2!} + \dots$$

C'est donc ce développement qu'il s'agit de calculer. Dans le quotient qui constitue la fonction y , on connaît le développement du numérateur λ^{x-1}

$$\lambda^{x-1} = \lambda^{-1} e^{x \log \lambda} = \lambda^{-1} \left(1 + \frac{x \log \lambda}{1} + \frac{x^2 \log^2 \lambda}{2!} + \dots \right)$$

Reste à calculer le développement de $\frac{1}{\Gamma(x)}$. Remarquons d'abord que l'on a $x \Gamma(x) = \Gamma(x+1)$.

D'où

$$\frac{1}{\Gamma(x)} = \frac{x}{\Gamma(x+1)}$$

Ainsi lorsque

$$x \rightarrow 0, \frac{1}{\Gamma(x)} \sim x;$$

on peut donc écrire

$$\frac{1}{\Gamma(x)} = x + \alpha x^2 + \beta x^3 + \dots$$

α, β, γ , etc. seront donnés par les dérivées successives de $\frac{1}{\Gamma(x)}$. Calculons la première

$$\frac{d}{dx} \frac{1}{\Gamma} = \frac{-\Gamma'}{\Gamma^2} = \frac{-\varphi(x)}{\Gamma(x)}$$

en posant $\varphi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$, dérivée logarithmique de $\Gamma(x)$.

Considérons à nouveau l'égalité : $x \Gamma(x) = \Gamma(x+1)$ et prenons la dérivée logarithmique ; il vient :

$$\frac{1}{x} + \frac{\Gamma'(x)}{\Gamma(x)} = \frac{\Gamma'(x+1)}{\Gamma(x+1)},$$

qui s'écrit aussi

$$\frac{1}{x} + \varphi(x) = \varphi(x+1)$$

Si $\varphi(x)$ ne peut être développé directement au voisinage de zéro, par contre $\varphi(x+1)$ peut l'être

$$\varphi(1+x) = \varphi(1) + \frac{x}{1!} \varphi'(1) + \frac{x^2}{2!} \varphi''(1) + \dots$$

Or les valeurs $\varphi(1), \varphi'(1), \varphi''(1)$, etc. (que nous écrirons par abréviation $\varphi_1, \varphi'_1, \varphi''_1, \dots$) peuvent être calculées en partant de la formule de Stirling qui donne

$$\log \Gamma(x) = \left(x - \frac{1}{2}\right) \log x - x + \frac{1}{2} \log 2\pi + \sum (-1)^{j-1} \frac{B_j}{2j(2j-1)x^{2j-1}}$$

et qui permet de calculer toutes les dérivées de $\log \Gamma(x)$.

Ainsi

$$\varphi(x) = \log x - \frac{1}{2x} + \sum (-1)^j \frac{B_j}{2^j x^{2j}} = \log x - \frac{1}{2x} - \frac{1}{12x^2} + \frac{1}{120x^4} - \frac{1}{252x^6} + \frac{1}{240x^8} - \frac{1}{132x^{10}} + \dots$$

$$\varphi'(x) = \frac{1}{x} + \frac{1}{2x^2} + \frac{1}{6x^3} - \frac{1}{30x^5} + \frac{1}{42x^7} - \frac{1}{30x^9} + \dots$$

$$\varphi''(x) = \frac{-1}{x^2} - \frac{1}{x^3} - \frac{1}{24x^4} + \frac{1}{6x^6} - \frac{1}{6x^8} + \frac{1}{10x^{10}} - \dots$$

$$\varphi'''(x) = \frac{2}{x^3} + \frac{3}{x^4} + \frac{2}{x^5} - \frac{1}{x^7} + \frac{4}{3x^9} - \frac{3}{x^{11}} + \dots$$

et en particulier leurs valeurs pour $x = 1$. Il faut bien se garder d'utiliser les formules qui précèdent en y faisant $x = 1$ car elles ne sont pas ou trop peu convergentes. Il est préférable de calculer $\varphi(x)$ et ses dérivées pour $x = 10$ (séries très convergentes) puis de revenir à 1 par les formules de récurrence

$$\varphi(1) = \varphi(10) - \frac{1}{9} - \frac{1}{8} \dots - 1$$

$$\varphi'(1) = \varphi'(10) + \frac{1}{81} + \frac{1}{64} + \dots + 1$$

$$\varphi''(1) = \varphi''(10) - \frac{2}{729} - \frac{2}{512} - \dots - 2$$

$$\varphi'''(1) = \varphi'''(10) + \frac{6}{6561} + \frac{6}{4096} + \dots + 6$$

On obtient ainsi très rapidement

$$\varphi_1 = -0,577215665$$

$$\varphi'_1 = +1,644934067$$

$$\varphi''_1 = -2,404113806$$

$$\varphi'''_1 = +6,493939402$$

D'où le développement de $\varphi(1+x)$, puis celui de $\varphi(x)$ qui s'écrit

$$\varphi(x) = -\frac{1}{x} + \varphi_1 + \varphi'_1 \frac{x}{1!} + \varphi''_1 \frac{x^2}{2!} + \dots$$

On a alors réuni les éléments nécessaires au calcul des coefficients $\alpha, \beta, \gamma, \dots$ du développement en série de Taylor de $\frac{1}{\Gamma(x)}$ en repartant de l'égalité

$$\frac{d}{dx} \frac{1}{\Gamma(x)} = -\varphi(x) \frac{1}{\Gamma(x)}$$

où l'on remplace $\varphi(x)$ et $\frac{1}{\Gamma(x)}$

par leurs développements :

$$1 + 2\alpha x + 3\beta x^2 + 4\gamma x^3 + \dots \equiv (x + \alpha x^2 + \beta x^3 + \dots) \left(\frac{1}{x} - \varphi_1 + \varphi_1' x + \varphi_1'' \frac{x^2}{2} + \dots \right)$$

$$\alpha = \varphi_1$$

$$\beta = \frac{\varphi_1^2 - \varphi_1'}{2}$$

$$\gamma = \frac{-\varphi_1^3 + 3\varphi_1\varphi_1' - \varphi_1''}{6}$$

$$\delta = \frac{\varphi_1^4 + 4\varphi_1\varphi_1'' - 6\varphi_1^2\varphi_1' + 3\varphi_1'^2 - \varphi_1'''}{24}$$

Enfin, le développement de $y(x)$ au voisinage de zéro s'obtient en multipliant celui de λ^{x-1} par celui de $\frac{1}{\Gamma(x)}$ ce qui donne :

$$y_0 + y_0' x + y_0'' \frac{x^2}{2} + \dots = (x + \alpha x^2 + \beta x^3 + \dots) \left(1 + \frac{x \log \lambda}{1} + \frac{x^2 \log^2 \lambda}{2} + \dots \right) e^{-\lambda} \lambda^{-\lambda}$$

d'où

$$y_0 = 0$$

$$y_0' = \lambda^{-1} e^{-\lambda}$$

$$y_0'' = 2 \lambda^{-1} e^{-\lambda} (\log \lambda - \varphi_1)$$

$$y_0''' = 3 \lambda^{-1} e^{-\lambda} [(\log \lambda - \varphi_1)^2 - \varphi_1']$$

$$y_0^{IV} = 4 \lambda^{-1} e^{-\lambda} [(\log \lambda - \varphi_1)^3 - 3(\log \lambda - \varphi_1)\varphi_1' - \varphi_1'']$$

$$y_0^V = 5 \lambda^{-1} e^{-\lambda} [(\log \lambda - \varphi_1)^4 - 6(\log \lambda - \varphi_1)^2\varphi_1' - 4(\log \lambda - \varphi_1)\varphi_1'' + 3\varphi_1'^2 - \varphi_1''']$$

et enfin

$$I(\lambda) = 1 + \lambda^{-1} e^{-\lambda} \left[\frac{1}{12} - \frac{(\log \lambda - \varphi_1)^2 - \varphi_1'}{240} + \frac{(\log \lambda - \varphi_1)^4 - 6(\log \lambda - \varphi_1)^2\varphi_1' - \dots}{6048} \dots \right]$$

D'où l'on peut tirer avec une bonne approximation la valeur de $I(\lambda)$ pour n'importe quelle valeur de λ .

Ces mêmes calculs permettent d'évaluer sans plus de difficultés les moments d'un ordre quelconque. Il suffit en effet de multiplier par x ou x^2 ou x^n le second membre de l'équation servant au calcul des dérivées. On a ainsi pour le moment du premier ordre (en nous bornant aux dérivées impaires) :

$$y_0' = 0$$

$$y_0''' = 6 \lambda^{-1} e^{-\lambda} (\log \lambda - \varphi_1)$$

$$y_0^V = 20 \lambda^{-1} e^{-\lambda} [(\log \lambda - \varphi_1)^3 - 3(\log \lambda - \varphi_1)\varphi_1' - \varphi_1'']$$

et pour le moment du second ordre :

$$y_0'' = 0$$

$$y_0^{IV} = 6 \lambda^{-1} e^{-\lambda}$$

$$y_0^V = 60 \lambda^{-1} e^{-\lambda} [(\log \lambda - \varphi_1)^2 - \varphi_1']$$

Nous avons indiqué précédemment comment on pouvait calculer les valeurs de $\varphi(x)$ en partant de la formule de Stirling. Cette détermination permet en opérant ensuite par interpolation de résoudre l'équation

$$\varphi(x) = \log \lambda$$

et par conséquent de calculer l'abscisse puis l'ordonnée du mode.

A cet effet, le tableau ci-dessous donne pour x variant de 0,05 en 0,05, entre 1 et 2, les valeurs de $\varphi(x)$ avec 8 décimales. Toutes les autres valeurs s'en déduisent par la formule

de récurrence
$$\varphi(x + 1) = \varphi(x) + \frac{1}{x}$$

Soit par exemple à calculer la valeur du mode correspondant à $\lambda = 3$ ou à $\log \lambda = 1,098612289$.

Dans l'intervalle (2, 3), $\varphi(x)$ sera compris entre 0,42278 et 0,92278; dans l'intervalle (3, 4) entre 0,92278 et 1,25611; la valeur cherchée est donc comprise dans cet intervalle. On trouve rapidement :

$$\varphi(3,45) = 1,0865020$$

$$\varphi(3,50) = 1,1031567$$

d'où

$$x_M = 3,4863 \dots$$

x	$\Gamma(x)$	$\varphi(x)$	x	$\Gamma(x)$	$\varphi(x)$
1,00	1,000 000 00	-0,577 215 66	1,55	0,888 868 35	+0,082 222 57
1,05	0,973 504 26	-0,497 844 99	1,60	0,893 515 35	+0,126 047 45
1,10	0,951 350 77	-0,423 754 94	1,65	0,900 116 82	+0,168 112 08
1,15	0,933 040 93	-0,354 326 68	1,70	0,908 638 74	+0,208 547 87
1,20	0,918 168 74	-0,289 039 89	1,75	0,919 062 53	+0,247 472 45
1,25	0,906 402 48	-0,227 453 53	1,80	0,931 383 77	+0,284 991 43
1,30	0,897 470 70	-0,169 190 89	1,85	0,945 611 18	+0,321 199 99
1,35	0,891 151 44	-0,113 928 01	1,90	0,961 765 83	+0,356 184 16
1,40	0,887 263 82	-0,061 384 54	1,95	0,979 880 65	+0,390 021 96
1,45	0,885 661 38	-0,011 316 42	2,00	1,000 000 00	+0,422 784 34
1,50	0,886 226 93	+0,036 489 97	2,05	1,022 179 48	+0,454 535 96