La longue genèse de la loi normale

Bernard Parzysz(*)

La loi normale est aujourd'hui devenue incontournable en statistique. Elle est née, au 18e siècle, de la recherche d'une « loi du hasard », et plus précisément du désir de certains « géomètres » de préciser la façon dont la fréquence se rapproche de la probabilité au cours de nombreuses répétitions d'une même épreuve aléatoire (en l'occurrence une épreuve de Bernoulli). Sur ce premier tronc est venue se greffer la question de la recherche d'une « loi des erreurs », cruciale pour tous ceux qui appliquaient les mathématiques à des domaines particuliers, notamment l'astronomie et la géodésie qui étaient en plein développement au « siècle des Lumières ». Je voudrais rappeler ici brièvement, sans m'attarder sur les aspects « techniques », quelques étapes de cette recherche, si fructueuse par ses conséquences.

1. Les expériences répétées

Après l'époque fondatrice du calcul des probabilités au milieu du 17e siècle (Pascal, Fermat, Huygens, Jacques Bernoulli, …), des scientifiques vont chercher à mieux cerner la notion de probabilité, dont l'existence a été postulée par Christiaan Huygens (1629-1695) dans *De ratiociniis in ludo aleae* (1657) : « *Quoique dans les jeux de hasard pur les résultats soient incertains, la chance qu'un joueur a de gagner ou de perdre a cependant une valeur déterminée.* » [Lanier 1992]. Leur but est de déterminer cette valeur de façon objective, notamment dans le cas d'une épreuve à deux issues. L'idée est que, sauf à fixer *a priori* par un moyen quelconque (comme par exemple un argument de symétrie), une valeur pour la chance que survienne l'une des issues (la probabilité de « succès »), on peut l'approcher de façon plus objective grâce à la répétition de l'expérience, et ce d'autant mieux que le nombre de répétitions sera grand, comme l'explique Jacques Bernoulli (1654-1705) dans son ouvrage posthume *Ars conjectandi* (1713) : « *Mais à la vérité s'offre ici à nous un autre chemin pour obtenir ce que nous cherchons. Ce qu'il n'est pas donné d'obtenir a priori l'est du moins a posteriori, c'est-à-dire qu'il sera possible de l'extraire en observant l'issue de nombreux exemples semblables, car on doit présumer que, par la suite, chaque fait peut arriver et ne pas arriver dans le même nombre de cas qu'il avait été constaté auparavant, dans un état de choses semblables, qu'il arrivait ou n'arrivait pas.* » [Bernoulli 1713]. Comme nous allons le voir, le but des mathématiciens qui se sont penchés sur cette question était, non seulement de prouver cette convergence – comme l'a fait Bernoulli –, mais aussi la préciser, c'est-à-dire de déterminer un majorant de l'erreur commise quand on évalue une probabilité inconnue par la fréquence observée après un grand nombre d'épreuves.

(*) Université d'Orléans et Laboratoire André-Revuz (Université Paris-Diderot).

1.1. Jacques Bernoulli

Dans l'*Ars conjectandi*, Bernoulli considère une épreuve aléatoire à deux issues [Henry 2004] : le tirage d'une boule dans une urne contenant t boules, dont r blanches et s noires ($r + s = t$) répété nt fois, avec remise (ce que nous appelons aujourd'hui un processus … de Bernoulli). On considère comme admis que, à chaque tirage, la probabilité de tirer une boule blanche de l'urne est égale à $p = \frac{r}{t}$. Bernoulli s'intéresse au nombre, B_{nt}, et à la fréquence, $\mathrm{F}_{nt} = \frac{\mathrm{B}_{nt}}{nt}$, de réalisations de l'une des deux issues (boule blanche). Il démontre alors que, quel que soit le nombre c (aussi grand que l'on veut), la probabilité que F_{nt} soit comprise entre $\frac{r-1}{t}$ et $\frac{r+1}{t}$ (valeurs qui encadrent la probabilité p à $\frac{1}{t}$ près) est supérieure à $\frac{c}{c+1}$ pour n suffisamment grand.

En langage moderne, cela signifie que la probabilité que F_{nt} s'écarte de p de plus de $\frac{1}{t}$ tend vers 0 lorsque n tend vers l'infini, et donc que F_{nt} est une approximation de p d'autant plus fiable que le nombre de répétitions de l'épreuve augmente.

B_{nt} suit la loi binomiale $\mathcal{B}(nt, p)$; la probabilité d'avoir k échecs(1) sur nt épreuves est donc $\mathrm{P}\left(\mathrm{B}_{nt} = nt - k\right) = \binom{nt}{k} p^{nt-k}(1-p)^k$. Bernoulli considère d'abord le développement du binôme dit « de Newton »(2) $(r+s)^{nt} = \sum_{k=0}^{nt} \binom{nt}{k} r^{nt-k} s^k$, constitué de $nt + 1$ termes que je noterai A(k). Il démontre alors que le plus grand terme de ce développement est $\mathrm{A}(ns) = \binom{nt}{ns} r^{nt-ns} s^{ns}$. Il considère ensuite les termes situés n rangs avant et après A(ns), soit A($ns - n$) et A($ns + n$) :

$$\mathrm{A}(ns-n) = \binom{nt}{ns-n} r^{nt-ns+n} s^{ns-n} = \mathrm{P}\left(\mathrm{B}_{nt} = nt - ns + n\right) \times t^{nt},$$

$$\mathrm{A}(ns+n) = \binom{nt}{ns+n} r^{nt-ns-n} s^{ns+n} = \mathrm{P}\left(\mathrm{B}_{nt} = nt - ns - n\right) \times t^{nt}.$$

Puis il montre que, pour n assez grand, le rapport de la somme des termes de rangs compris entre $ns - n$ et $ns + n$ à la somme de tous les autres termes peut être rendu

(1) Bernoulli s'intéresse en effet au nombre d'échecs et non au nombre de succès.
(2) Mais déjà mentionné en 1593 dans l'ouvrage chinois *Souan Fa Long Tsong* (Principes du calcul) [Marre 1846].

plus grand que tout nombre c donné (fig. 1).

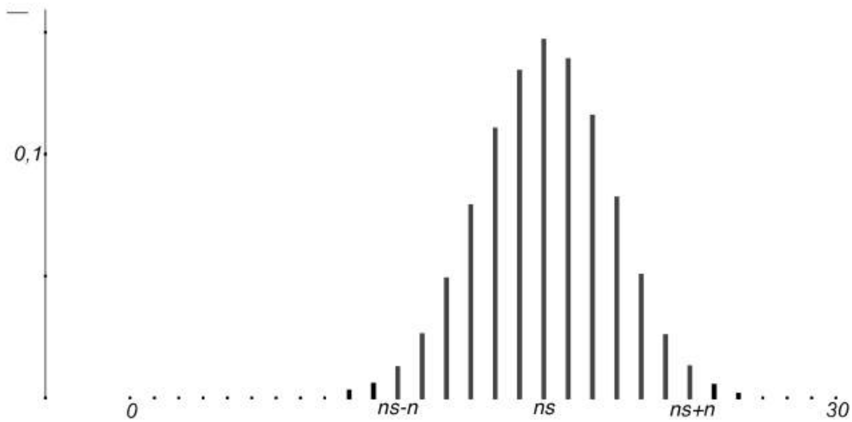

Figure 1
(avec $r = 2$, $s = 3$ (d'où $t = 5$) et $n = 6$)

Il en résulte que le rapport de la somme des termes de rang compris entre $ns - n$ et $ns + n$ à la somme t^{nt} de tous les termes (c'est-à-dire la probabilité que B_{nt} soit compris entre $nt - ns - n$ et $nt - ns + n$) est supérieur à $\frac{c}{c+1}$ (qui a pour limite 1 quand c tend vers $+\infty$). On a donc finalement :

$$P(nt - ns - n \leq B_{nt} \leq nt - ns + n) > \frac{c}{c+1},$$

$$P(n(r-1) \leq B_{nt} \leq n(r+1)) > \frac{c}{c+1}$$

ou enfin, en quantifiant cette inégalité :

$$\forall t > 0\ \forall c > 0\ \exists N\ \forall n > N\ P(n(r-1) \leq B_{nt} \leq n(r+1)) > \frac{c}{c+1},$$

$$P\left(\frac{r-1}{t} \leq F_{nt} \leq \frac{r+1}{t}\right) > \frac{c}{c+1}.$$

D'où :

$$\forall t > 0\ \lim_{n\to+\infty}\left[P\left(|F_{nt} - p| \leq \frac{1}{t}\right)\right] = 1$$

(convergence en probabilité de F_{nt} vers p).

1.2. Abraham de Moivre

Bernoulli montre donc que la probabilité que la fréquence de succès ne s'écarte pas de plus d'une quantité donnée de la probabilité de succès tend vers 1 lorsque le nombre d'épreuves devient très grand, mais il ne réussit pas à donner une approximation de cette probabilité pour n (assez grand) donné(3). C'est ce à quoi va immédiatement s'attaquer Abraham de Moivre (1667-1754). Dans sa *Doctrine of chances* (1ère édition 1718, 3e éd. 1756) [Lanier et Trotoux 1995], il reprend l'idée de Bernoulli de considérer les termes symétriques par rapport au plus grand terme du développement du binôme et se place d'abord dans le cas particulier où les deux issues de l'expérience, répétée n fois, sont également probables (comme à pile ou face avec une pièce équilibrée). La probabilité d'obtenir k fois l'une des deux issues $(0 \le k \le n)$ est donc $\dfrac{\binom{n}{k}}{2^n}$. Moivre commence par donner, à l'aide de la formule que – écrit-il – lui a communiquée son ami James Stirling (c'est-à-dire $n! \sim \sqrt{2\pi n}\left(\dfrac{n}{\mathrm{e}}\right)^n$), un équivalent à l'infini du terme central du développement de

$$\left(\frac{1}{2}+\frac{1}{2}\right)^n = \sum_{k=0}^{n} \frac{\binom{n}{k}}{2^n}.$$

Soit, en posant $m = n/2$ (et en supposant implicitement n pair) :

$$\frac{\binom{n}{[m]}}{2^n} \sim \frac{2}{\sqrt{2\pi n}}$$

($[m]$ désignant la partie entière de m). Puis il donne un équivalent du logarithme du rapport de ce terme central à ceux qui en sont distants de l (en supposant implicitement l négligeable devant n) :

$$\ln \frac{\binom{n}{m}}{\binom{n}{m \pm 1}} \sim \left(m+l-\frac{1}{2}\right)\ln(m+l-1)+\left(m-l+\frac{1}{2}\right)\ln(m-l+1)-2m\ln m+\ln\frac{m+1}{[m]}.$$

Il aboutit alors (corollaire 1) à l'équivalent suivant :

$$\frac{\binom{n}{m-l}}{2^n} \sim \exp\left(-\frac{2l^2}{n}\right) \times \frac{2}{\sqrt{2\pi n}}$$

où l'on voit pointer le bout du nez de la future loi normale, équivalent qu'il utilise pour sommer les termes dont les rangs sont compris entre $m - l$ et $m + l$.

Enfin, en posant $l = s\sqrt{n}$, il trouve comme équivalent de cette somme le produit de $\dfrac{2}{\sqrt{2\pi}}$ par la somme de la série de terme général $\dfrac{(-2)^k s^{2k+1}}{k!(2k+1)}$.. Il l'utilise avec $s = \dfrac{1}{2}$

(3) À cause des majorations trop « brutales » qu'il utilise.

dans le but d'obtenir une valeur numérique approchée de cette somme, et conclut (corollaire 3) :
« *la probabilité qu'un événement qui a un nombre égal de chances de se produire ou non, n'ait lieu ni plus que* $\frac{n}{2}+\frac{\sqrt{n}}{2}$ *fois, ni moins que* $\frac{n}{2}-\frac{\sqrt{n}}{2}$ *fois, sera exprimée par (…) 0,682688.* ». Moivre a donc quantifié la probabilité que la fréquence de succès soit comprise entre les valeurs $\frac{1}{2}-\frac{1}{2\sqrt{n}}$ et $\frac{1}{2}+\frac{1}{2\sqrt{n}}$ puisqu'il fournit, dans ce cas particulier, une approximation de la probabilité que la fréquence observée se trouve dans un intervalle donné centré sur la probabilité de succès. Plus précisément – et pour parler comme aujourd'hui – il donne, pour le jeu de pile ou face, l'intervalle de fluctuation $\left[\frac{1}{2}-\frac{1}{2\sqrt{n}};\frac{1}{2}+\frac{1}{2\sqrt{n}}\right]$ au seuil de 68%.

Moivre étend ensuite ses résultats au cas où il n'y a pas équiprobabilité, et conclut (corollaire 10) : « *Si les probabilités de succès et d'échec sont dans un rapport donné d'inégalité, les problèmes relatifs à la somme des termes du binôme* $(a+b)^n$ *seront résolus avec la même facilité que ceux dans lesquels les probabilités de succès et d'échec sont dans un rapport d'égalité.* ».

N.B. En 1770, s'attaquant au même problème que son oncle Jacques, Daniel Bernoulli (1700-1782) tombera, lui aussi, sur une fonction du type $x \to \mathrm{e}^{-kx^2}$, et il en donnera une table succincte.

1.3. Pierre-Simon Laplace

Dans sa *Théorie analytique des probabilités* (1812), Pierre-Simon Laplace (1749-1827) continue à creuser le sillon tracé par Bernoulli et Moivre. Il considère d'abord le développement de $\left(p+(1-p)\right)^{x+x'}$ (avec $x + x' = n$) et montre que son plus grand terme s'obtient, pour x et x' très grands, lorsque $\frac{x}{x'}$ est voisin de $\frac{p}{1-p}$. Comme Moivre, il s'intéresse ensuite aux termes situés l rangs après et l rangs avant ce terme maximum (l étant au plus de l'ordre de $\sqrt{n}$), afin d'évaluer la somme des termes dont les rangs sont compris entre les deux. Pour cela, il utilise, outre les formules de Stirling et de Taylor, une formule qui ramène la somme d'une série à celle d'une intégrale. En posant $\mathrm{T}=\frac{l\sqrt{n}}{\sqrt{2xx'}}$ il trouve un intervalle de fluctuation tel que :

$$\mathrm{P}\left[-\frac{\mathrm{T}\sqrt{2xx'}}{n\sqrt{n}}\le\frac{\mathrm{X}}{n}-p\le\frac{\mathrm{T}\sqrt{2xx'}}{n\sqrt{n}}\right]=\frac{2}{\sqrt{\pi}}\int_0^{\frac{\mathrm{T}\sqrt{n}}{\sqrt{2xx'}}}\mathrm{e}^{-t^2}\,dt+\frac{\sqrt{n}}{\sqrt{\pi}\sqrt{2xx'}}\mathrm{e}^{-\mathrm{T}^2}$$

(où X désigne le nombre de succès et T un réel strictement positif).

Puis il termine en « retournant » la formule et en donnant un intervalle de confiance pour p :

$$\mathrm{P}\left[f_n - \frac{\mathrm{T}}{\sqrt{n}}\sqrt{2f_n(1-f_n)} \leq p \leq f_n + \frac{\mathrm{T}}{\sqrt{n}}\sqrt{2f_n(1-f_n)}\right]$$
$$= \frac{2}{\sqrt{\pi}}\int_0^{\mathrm{T}} \mathrm{e}^{-t^2}\,dt + \frac{1}{\sqrt{n\pi}} \cdot \frac{\mathrm{e}^{-\mathrm{T}^2}}{\sqrt{2f_n(1-f_n)}}$$

où f_n est la fréquence de succès susceptible d'être observée sur n épreuves (n suffisamment grand).

On a donc ici un moyen d'estimer la probabilité, non seulement que la fréquence observable se situe dans un intervalle donné centré sur la probabilité de succès, mais également que la probabilité de succès se situe dans un intervalle donné centré sur la fréquence observable. D'autre part, on y voit apparaître l'intégrale de e^{-t^2}, dont on n'a hélas pas de primitive exprimable à l'aide des fonctions usuelles. Mais il est néanmoins possible de l'approcher, à l'instar de Daniel Bernoulli (voir plus haut).

2. La théorie des erreurs

Les lois physiques sont essentiellement issues d'invariants identifiés à partir d'observations et de mesures. De tout temps on avait bien sûr constaté que plusieurs mesures d'une même quantité donnaient en général des résultats différents, quoique voisins (sauf erreur de mesure ou de méthode), et la question était de déterminer, sinon la « vraie » valeur de cette quantité, du moins la « meilleure » valeur (c'est le problème de l'*estimation*).

C'est le cas en particulier pour l'astronomie et la géodésie, sciences en plein développement au siècle des Lumières et dont les méthodes de mesures de distances et d'angles étaient très raffinées [Guedj 2000].

L'idée de base est donc de substituer à la série de mesures $(x_1, x_2, \ldots, x_n)$ une série $(\mu, \mu, \ldots, \mu)$, le nombre μ étant choisi de façon à minimiser une certaine fonction des mesures $f(x_1, x_2, \ldots, x_n, \mu)$ [Armatte 2004].

2.1. Les écarts

Au milieu du 18e siècle on va commencer à s'intéresser, non plus aux valeurs, mais aux écarts à la vraie valeur (inconnue), c'est-à-dire aux variables $e_i = x_i - x$.

En 1756-57, Thomas Simpson (1710-1761) fait la remarque que ce ne sont pas les observations qui importent, mais les écarts entre ces observations et la vraie valeur (inconnue). Il fait alors une analyse probabiliste de ces écarts (qu'il appelle erreurs, d'où le nom de « théorie des erreurs ») et suppose que les erreurs positives et négatives également distantes de la « vraie » valeur ont la même probabilité de survenir (symétrie de la distribution). Il étudie en particulier le cas d'une distribution continue rectangulaire (répartition uniforme des erreurs) et celui d'une distribution triangulaire (les erreurs décroissent linéairement : les grandes erreurs ont donc moins de chance d'apparaître que les petites). Il montre que l'erreur sur la moyenne de plusieurs observations est inférieure à celle commise sur chaque observation prise séparément. (certains pensaient alors que le mieux était de faire une seule mesure, mais en y mettant tout le soin possible).

Jean III Bernoulli (1744-1807) fait une remarque analogue vers 1770, dans l'article « milieu » du Supplément à l'Encyclopédie : « *en prenant simplement le milieu arithmétique, on ne tient pas compte du plus ou moins de probabilité de l'exactitude des observations* » (cité dans [Boyé & Comairas 2002], p. 31).

Il s'agit maintenant d'estimer « au mieux » la vraie valeur inconnue.

Parallèlement, en 1757, Ruđer Bošcović (1711-1787) écrit que la meilleure estimation est celle pour laquelle :

– la somme des écarts positifs est égale à celle des écarts négatifs,
– la somme des valeurs absolues des écarts est aussi petite que possible.

La fonction *f* du début est donc ici $f\left(x_1,x_2,\ldots,x_n,x\right)=\sum_{i=1}^{n}\left|x_i-x\right|$.

Dans ce cas, on démontre que la « meilleure » valeur n'est pas la moyenne, mais la médiane (on trouvera une idée de démonstration de ce résultat à l'adresse : http://euler.acversailles.fr/wm3/pi2/co_moyenne_mediane/index.jsp).

Mais en 1805, dans ses *Nouvelles méthodes pour la détermination des orbites des comètes*, Adrien-Marie Legendre (1752-1833) écrit : « *De tous les principes que l'on peut proposer pour cet objet, je pense qu'il n'en est pas de plus général, de plus exact, ni d'une application plus facile que celui dont nous avons fait usage dans les recherches précédentes, et qui consiste à rendre minimum la somme des carrés des erreurs.* »

La fonction *f* mentionnée plus haut est donc ici $f\left(x_1,x_2,\ldots,x_n,x\right)=\sum_{i=1}^{n}\left(x_i-x\right)^2$.

Dans ce cas, la « meilleure » valeur est la moyenne arithmétique. En effet, soit, pour x réel, $f\left(x\right)=\sum_{i=1}^{n}\left(x_i-x\right)^2$. On a $f'(x) = 2n(x - m)$, où m est la moyenne arithmétique de la série, et f admet donc un minimum en m.

Or, la moyenne arithmétique est justement l'estimation qu'on utilisait intuitivement jusque-là.

2.2. Les lois de densité

Il s'agit maintenant de savoir comment se répartissent les erreurs par rapport à la « vraie » valeur (inconnue). Deux principes – de bon sens – semblent acquis :

– on a autant de chances d'avoir une erreur négative qu'une erreur positive ;
– on a moins de chance d'avoir une grosse erreur qu'une petite.

C’est-à-dire que la loi aura *grosso modo* une allure « en cloche ».

Diverses « lois de facilité » des erreurs (nous dirions aujourd’hui « courbes de densité ») sont alors proposées. Ainsi, Jean-Henri Lambert (1728-1777) propose en 1760 la courbe suivante (fig. 2) :

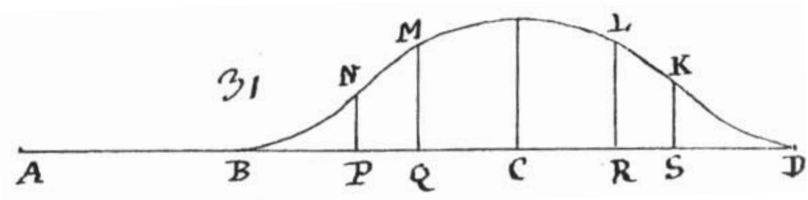

Figure 2

Il ne dit pas comment il l’a obtenue, mais on pense que c’est par lissage d’un histogramme (borné) correspondant à de nombreuses expériences.

En 1773, Joseph-Louis Lagrange (1736-1813) indique qu’« *il n’y a aucun doute que les petites erreurs ont lieu plus souvent que les grandes* ».

Mais tout le monde considère que les erreurs sont bornées : une valeur trop éloignée des autres sera éliminée.

En 1774, Laplace indique quelques conditions que doit vérifier la loi de facilité des erreurs :
« *La loi suivant laquelle cette vraisemblance* [= probabilité d’apparition] *diminue à mesure que l’observation s’éloigne de la vérité nous est inconnue. Supposons donc (fig. 2) que le point V soit le véritable instant du phénomène* [il s’agit d’une mesure de temps] *(…) en nommant x l’abscisse VP, et y l’ordonnée correspondante PM, nous représenterons l’équation par celle-ci : $y = \varphi(x)$.*

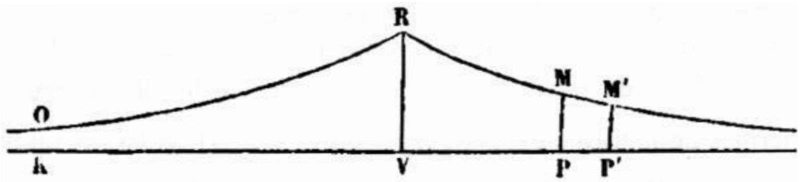

Figure 3

Or voici les propriétés de cette courbe [fig. 3] :
1° elle doit être partagée en deux parties entièrement semblables par la droite VR, car il est tout aussi probable que l’observation s’écartera de la vérité à droite comme à gauche ;
2° elle doit avoir pour asymptote la ligne KP, parce que la probabilité que l’observation s’éloigne de la vérité à une distance infinie est nulle ;
3° l’aire entière de cette courbe doit être égale à l’unité, puisqu’il est certain que l’observation tombera sur un des points de la droite KP. »

Laplace étend l'intervalle des possibles à $\mathbb{R}$ tout entier et retient le principe des moindres erreurs de Bošcović, c'est-à-dire qu'il centre la courbe en « *l'instant tel qu'en le prenant pour milieu, la somme des erreurs à craindre* [en valeur absolue], *multipliées par leur probabilité, soit un minimum* ».

Il ajoute une hypothèse supplémentaire : « *nous n'avons aucune raison de supposer une autre loi aux ordonnées qu'à leurs différences* », c'est-à-dire que les ordonnées suivent la même loi que leurs différences, soit $dy/y = \mathrm{C^{te}}$ (la constante étant ici négative, puisque « *les petites erreurs ont lieu plus souvent que les grandes* »). Mais il ne donne aucune véritable justification théorique de cette dernière condition.

Il déduit de l'équation différentielle que la loi des erreurs est une loi exponentielle bilatère, du type $y = \lambda e^{-k|x|}$, où k est une constante et λ un coefficient déterminé en fonction de k.

On doit en outre avoir (symétrie) $\int_0^{+\infty} \lambda e^{-kx}\, dx = 0{,}5$, soit $\lambda/k = 0{,}5$, d'où $\lambda = k/2$.

La fonction est donc définie par $y = \frac{k}{2} e^{-k|x|}$ (appelée depuis *première loi de Laplace*).

Néanmoins, en 1785 Laplace écrit :

« *L'intégrale* $\int dt \cdot e^{-t^2}$ *se rencontre fréquemment dans cette analyse et, par cette raison, il serait très utile de former une Table de ses valeurs, depuis* $t = \infty$ *jusqu'à* $t = 0$. »

2.3. Carl-Friedrich Gauss

L'idée des moindres carrés est ensuite reprise en 1809 par Gauss (1777-1855), qui montre que, si la moyenne des erreurs (indépendantes) trouvée par la méthode des moindres carrés est la valeur la plus probable, alors la loi des erreurs est d'un certain type (il s'agit de la future loi normale) [Lefort 2012]:

Soit la série d'écarts $e_i = x_i - x$ qui :

– sont mutuellement indépendantes (on recommence la procédure pour faire une nouvelle mesure) ;
– suivent tous la même loi continue de densité φ (il s'agit des écarts relatifs à des mesures d'un même objet faites dans les mêmes conditions).

Pour x donné, la probabilité d'obtenir des écarts compris entre e_i et $e_i + de_i$ est alors

$$\left(\prod_{i=1}^{n} \varphi(e_i)\right) de_1 \ldots de_n.$$

Inversement, on peut chercher une valeur μ pour laquelle la probabilité d'obtenir le n-uplet des valeurs observées $(x_1, \ldots, x_n)$ est maximum. Pour μ on a donc un maximum de l'expression $\prod_{i=1}^{n} \varphi(e_i)$, et par conséquent un maximum de son logarithme, soit $\psi(x) = \sum_{i=1}^{n} \ln\big(\varphi(x_i - x)\big)$. Sa dérivée $\psi'(x) = \sum_{i=1}^{n} \frac{d\big(\ln\big(\varphi(x_i - x)\big)\big)}{dx}$ s'annule donc en μ.

On veut aussi que μ soit égal à la moyenne arithmétique m de la série qui par définition vérifie $\sum_{i=1}^{n} (x_i - x) = 0$. Gauss en déduit qu'il suffit pour cela d'avoir l'égalité terme à terme (à un coefficient près), soit, pour tout i :

$$\frac{d\big(\ln\big(\varphi(x_i - x)\big)\big)}{dx} = k(x_i - x).$$

D'où, en intégrant : $\ln\big(\varphi(x_i - x)\big) = -\frac{k}{2}(x_i - x)^2 + \mathrm{C}^{te}$, soit

$$\varphi(x_i - x) = \mathrm{C}\exp\left(-\frac{k}{2}(x_i - x)^2\right).$$

On vérifie qu'avec cette fonction φ on a $\prod_{i=1}^{n} \varphi(e_i) = \mathrm{C}\exp\left(-\frac{k}{2}\sum_{i=1}^{n} e_i^2\right)$, qui est maximum lorsque $\sum_{i=1}^{n} e_i^2$ est minimum, c'est-à-dire quand x est la moyenne arithmétique des mesures.

Ainsi donc, puisque la « vraie valeur » de la méthode des moindres carrés n'est autre que la moyenne arithmétique m des observations, Gauss obtient une « loi des erreurs » qui s'exprime sous la forme $f : x \to \mathrm{C}e^{-\frac{k(x-m)^2}{2}}$, où C et k sont des constantes respectivement égales à $\frac{1}{\sigma\sqrt{2\pi}}$ et $\frac{1}{\sigma^2}$ (σ^2 étant la variance des observations)(4).

3. Et ensuite ?

L'astronome belge Adolphe Quételet (1796-1855) a ensuite trouvé la loi normale dans d'autres phénomènes que les erreurs de mesure, et notamment dans des séries biologiques (tour de taille des soldats), où il interprétait les écarts à la moyenne comme des erreurs (!) par rapport à un idéal. C'est Francis Galton (1822-1911) qui, en 1889, a le premier parlé de « courbe normale » pour désigner la distribution des

(4) Ces valeurs s'obtiennent à partir de l'intégrale de Gauss ($\int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}} dt = \sqrt{2\pi}$), ainsi que du changement de variable $x = \sigma t + m$.

erreurs de mesure, distribution qu'il a également appliquée à des données biologiques. L'école française, elle, préférait appeler cette loi « deuxième loi de Laplace » ou « loi de Laplace-Gauss ». C'est en fait Karl Pearson (1857-1936) qui a popularisé le terme de « loi normale » à partir de 1893. Mais il a finalement reconnu (en 1920) que ce terme était mal choisi, car il avait une connotation péjorative pour les autres lois. Enfin, Jacques Hadamard (1865-1963) montre l'intérêt de la loi normale dans le contrôle des produits industriels et Émile Borel (1877-1956) précise les conditions générales de réalisation d'une distribution normale : la distribution d'une variable X qui dépend

- d'une part d'une cause prépondérante constante
- d'autre part de nombreuses causes secondaires ayant des effets aléatoires petits, additifs, symétriques, et indépendants

tend vers la loi normale, et ceci d'autant plus que le nombre de causes secondaires est grand(5).

C'est le cas pour de nombreux phénomènes, relevant de la physique, de la biologie, de l'industrie, etc. Mais l'omniprésence de la loi normale ne doit pas nous faire oublier l'existence d'autres types de distribution, quitte à tempérer l'enthousiasme de Galton.

En conclusion de ce survol rapide de la (lente) genèse de la loi normale, on a pu voir qu'elle apparaît historiquement comme limite d'une distribution binomiale, lorsque le nombre de répétitions de l'épreuve de Bernoulli devient très grand, ainsi que l'indique le théorème de Moivre-Laplace qui figure au programme 2012 de Terminale S. :

> *Pour introduire la loi normale* $\mathcal{N}(0,\ 1)$ *on s'appuie sur l'observation des représentations graphiques de la loi de la variable aléatoire* $\mathrm{Z}_n = \dfrac{\mathrm{X}_n - np}{\sqrt{np(1-p)}}$
>
> *où* X_n *suit la loi binomiale* $\mathcal{B}(n, p)$*, et cela pour de grandes valeurs de n et une valeur de p fixée entre 0 et 1. Le théorème de Moivre-Laplace assure que pour tous réels a et b,* $\mathrm{P}(\mathrm{Z}_n \in [a,\ b])$ *tend vers* $\displaystyle\int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$ *lorsque n tend vers* $+\infty$.

La difficulté majeure réside bien sûr dans le passage du discret au continu(6). Et, finalement, en référence au nouveau programme de Terminale S, la meilleure

(5) Le théorème central limite montre que la moyenne réduite de variables indépendantes de même loi converge vers la normale centrée réduite. Ceci explique que la somme d'un grand nombre de telles variables infinitésimales, qui s'additionnent pour engendrer une erreur aléatoire dans une mesure, suive une loi « presque » normale. Il en va d'ailleurs de même pour des variables n'ayant pas même loi mais qui ont des amplitudes analogues.

(6) Une fois introduite la loi normale comme le programme le préconise, on pourra réaliser la contre-épreuve consistant à « discrétiser celle-ci pour faire voir que les diagrammes en bâtons de la binomiale et de la « normale discrétisée » de mêmes paramètres se superposent d'autant mieux que n est grand et p proche de 0,5 [Parzysz 2007].

approximation d'une loi binomiale n'est-elle pas… cette loi binomiale elle-même ? La puissance des outils actuels de calcul permet en effet, dans la grande majorité des cas, de travailler directement sur la « vraie » loi et non par le truchement d'un substitut, aussi intéressant soit-il(7).

Bibliographie

Armatte, M. (2004). La théorie des erreurs (1750-1820), enjeux, problématiques, résultats. *Histoires de probabilités et de statistiques* (coord. par E. Barbin et J.-P. Lamarche), p. 141-160. Ed. Ellipses.

Bernoulli, J. (1713). *Ars conjectandi*. 4ème partie. Traduit du latin par N. Meusnier. Ed. IREM de Rouen 1987.

Boyé, A. & Comairas, M.-C. (2002). Moyenne, médiane, écart-type. *Repères-IREM* 48, 27-40.

Dutarte, P. & Kern, C. (2002). *La statistique inférentielle en quatre séances*. Carnet de stage. Ed. IREM Paris-Nord.

Guedj, D. (2000). *Le mètre du monde*. Ed. du Seuil.

Henry, M. (2004). La démonstration par Jacques Bernoulli de son théorème. *Histoires de probabilités et de statistiques*, p. 121-140. Ed. Ellipses.

Lanier, D. (1992). Huygens, l'espérance et l'infini. *Histoires d'infini* (Actes du 9ème colloque inter-IREM Épistémologie et Histoire des Mathématiques). Ed. IREM de Brest, p. 555-577.

Lanier, D. & Trotoux, D. (1995). La loi des grands nombres, le théorème de de Moivre-Laplace. *Contribution à une approche statistique de l'enseignement des mathématiques*, p. 259-294. Ed. Presses universitaires de Franche-Comté. http://www.math.ens.fr/culturemath/histoire%20des%20maths/pdf/LoidesGrandsNombres.pdf

Lefort, X. (2012). Introduction de la loi normale à partir du texte original de Gauss. *Les mathématiques éclairées par l'histoire* (coordonné par E. Barbin), p. 115-130. Ed. Vuibert

Marre, A. (1846). Du binôme de Newton, antérieurement à Newton. *Nouvelles annales de mathématiques*. Première série, tome 5, p. 488-496.

Parzysz, B. (2007). Loi binomiale, courbe en cloche et tableur. *Bulletin de l'APMEP* 473, 880-886.

Pichard, J.-F. (2005). Théorie des erreurs, courbes en cloche et normalité. *Statistique au lycée* (coord. par B. Chaput et M. Henry), p. 219-246. Ed. APMEP.

(7) Merci à Michel Henry pour son amicale et attentive relecture.