REVUE DE STATISTIQUE APPLIQUÉE

J. COULLET

Méthode de calcul des fonctions de répartition des lois usuelles de probabilité par le calcul direct des intégrales

Revue de statistique appliquée, tome 36, n° 1 (1988), p. 5-18

<http://www.numdam.org/item?id=RSA_1988__36_1_5_0>

© Société française de statistique, 1988, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) 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/

Rev. Statistique Appliquée, 1988, **XXXVI** (1), 5-18

MÉTHODE DE CALCUL DES FONCTIONS DE RÉPARTITION DES LOIS USUELLES DE PROBABILITÉ PAR LE CALCUL DIRECT DES INTÉGRALES

J. COULLET

Directeur du Département Recherche Commerciale et Prévisions à la S.E.I.T.A.

I. Principe général de la méthode

Bien que des tables statistiques relatives aux fonctions de répartition des distributions probabilistes, aussi bien les plus courantes que les moins utilisées, existent depuis de nombreuses années, l'application des méthodes de calcul automatique à ces fonctions reste d'un intérêt toujours actuel. En effet, le calcul automatique allie les qualités de précision, de rapidité et de commodité. Il permet d'éviter les calculs d'interpolation de tables, souvent fastidieux et peu précis, en particulier dans le cas de lois à un ou plusieurs paramètres. Au surplus, les routines de calcul automatique peuvent être intégrées dans des processus plus généraux, comme les processus de contrôle.

On sait que, le plus souvent, les fonctions de répartition sont définies par des intégrales irréductibles. La méthode la plus utilisée pour les calculer consiste à élaborer une fonction d'approximation ajustant la fonction de répartition. De nombreux algorithmes de ce type ont été mis au point et programmés [1] et [3] (références citées à titre d'exemple).

On va exposer ci-après une contribution à ces calculs procédant d'une approche différente, à savoir le calcul direct des intégrales. Cette méthode constitue d'un certain point de vue un retour aux sources, car c'est ainsi que furent obtenues les premières tables, dont certaines sont très précises, comme la table de la fonction Bêta incomplète de Karl PEARSON. Toutefois, si l'idée est simple, sa mise en application se heurte à des difficultées pratiques de réalisation souvent très contraignantes. Au surplus, pour valider la précision des résultats, certains calculs complexes (qui ne sont d'ailleurs pas sans intérêt du point de vue probabiliste et statistique) sont nécessaires.

II. Distributions concernées

Cette programmation a été appliquée aux lois Gamma et Bêta. Ce choix n'est pas neutre : on sait en effet que de nombreuses distributions (χ^2, loi normale, Poisson, Fisher-Snedecor, Student-Fisher, loi binomiale) dérivent de ces lois, ainsi que le montrent les tableaux ci-après.

Mots clés : *Distribution Gamma, Distribution Bêta, Intégrale de Boole, Formule de Stirling.*

Notations utilisées
(x désigne toujours la variable, discrète ou continue)

Loi	Fonction de répartition
Loi Gamma de paramètre p	$F_{GA}(p\ ;\ x)$
Loi de χ^2 de ddl ν	$F_{\chi^2}(\nu\ ;\ x)$
Loi normale réduite	$F_{NO}(x)$
Loi de Poisson de paramètre m	$F_{PO}(m\ ;\ x)$
Loi Bêta de paramètre p et q	$F_{BE}(p\ ,\ q\ ;\ x)$
Loi de Fisher-Snédecor de ddl ν_1 et ν_2	$F_{FS}(\nu_1\ ,\ \nu_2\ ;\ x)$
Loi de Student-Fisher de ddl ν	$F_{SF}(\nu\ ;\ x)$
Loi Binomiale de paramètres p et n	$F_{BI}(p\ ,\ n\ ;\ x)$

Identités d'équivalence

$$F_{\chi^2}(\nu\ ;\ x) = F_{GA}\left(\frac{\nu}{2}\ ;\ \frac{x}{2}\right)$$

$$\begin{cases} x \geqslant 0 & F_{NO}(x) = \dfrac{1}{2}\left[1 + F_{GA}\left(0{,}5\ ;\ \dfrac{x^2}{2}\right)\right] \\ x < 0 & F_{NO}(x) = \dfrac{1}{2}\left[1 - F_{GA}\left(0{,}5\ ;\ \dfrac{x^2}{2}\right)\right] \end{cases}$$

$$F_{PO}(m\ ;\ x) = 1 - F_{GA}(x + 1\ ;\ m)$$

$$F_{FS}(\nu_1\ ,\ \nu_2\ ;\ x) = F_{BE}\left(\frac{\nu_1}{2}\ ,\ \frac{\nu_2}{2}\ ;\ \frac{\nu_1 x}{\nu_1 x + \nu_2}\right)$$

$$\begin{cases} x \geqslant 0 & F_{SF}(\nu\ ;\ x) = \dfrac{1}{2}\left[1 + F_{BE}\left(0{,}5\ ,\ \dfrac{\nu}{2}\ ;\ \dfrac{x^2}{x^2 + \nu}\right)\right] \\ x < 0 & F_{SF}(\nu\ ;\ x) = \dfrac{1}{2}\left[1 - F_{BE}\left(0{,}5\ ,\ \dfrac{\nu}{2}\ ;\ \dfrac{x^2}{x^2 + \nu}\right)\right] \end{cases}$$

$$F_{BI}(p\ ,\ n\ ;\ x) = F_{BE}[(n - x)\ ,\ (x + 1)\ ;\ (1 - p)]$$

III. Performances du logiciel mis au point

Le programme calcule la fonction de répartition des fonctions suivantes : Lois Gamma, χ^2, Normale, Poisson, Bêta, Fisher-Snédecor, Student-Fisher, Binomiale. La programmation de ces lois a été effectuée à partir des lois Bêta et Gamma.

L'édition peut être faite, soit par valeurs individuelles, soit par tableaux.

Le programme donne, en virgule flottante, avec la précision disponible, les valeurs de F et $1 - F$, ce qui, pour les queues de distribution, est nettement plus avantageux que de donner F seulement (considérer, par exemple le cas : F = 0,9999993).

Autres caractéristiques :

Loi Gamma

— Domaine de variation de p : 0,3 ; 1 000.000.

— Précision : 10^{-7} (7 chiffres exacts en format virgule flottante pour F et $1-F$). La précision est choisie par l'utilisateur entre 10^{-4} et 10^{-7}. Les probabilités inférieures à 10^{-20} ne sont pas calculées.

— Temps de calcul : De 3 à 18 secondes, moyenne 9s (valeurs approximatives), pour une précision de 7 chiffres. Le temps diminue si la précision demandée est plus faible.

La limite de 0,3 pour p s'explique par le fait que nous avons voulu traiter le cas $p = 0{,}5$, utile pour la loi normale et que, pour p voisin de zéro, la convergence des sommations utilisées pour le calcul de l'intégrale n'est pas assurée.

Loi Bêta

— Domaine de variation de p et q : 0,3 ; 1.000 000.
— Précision : entre 10^{-7} et 10^{-5} pour F et $1-F$.
— Temps de calcul : analogue à celui de la loi Gamma.
— La valeur p (resp q) = 0,5 est utilisée dans le cas de la loi de Student-Fisher.
— Pour les petites valeurs de p (resp q) on rencontre également des problèmes de convergence des sommations utilisées. C'est pour cela que nous avons limité p (resp q) à 0,3.

Pour les autres lois se reporter aux identités d'équivalence précédemment données.

Les caractéristiques de la calculatrice utilisée sont les suivantes :

— Matériel : Micro-ordinateur HP 85,
— Langage de programmation : BASIC,
— Format des nombres : 12 chiffres en virgule flottante, ($10^{-499} < |x| < 10^{499}$),
— Les fonctions usuelles, en particulier le logarithme et l'exponentielle, sont disponibles avec la précision standard,
— Mémoire vive : 30 KO,
— Mémoire de stockage : cassette de 200 KO.

Le programme occupe un volume de 17 KO.

Les performances peuvent être améliorées par l'emploi de systèmes plus efficaces. Par exemple, par une traduction en langage QUICK BASIC compilé, actuellement disponible et compatible IBM (16 bits, MS DOS), la précision de 10^{-7} serait accessible partout, et le temps de calcul serait divisé par 3 (moyenne 3 secondes).

En résumé, on peut dire qu'en ce qui concerne la précision des calculs, les performances de cette méthode sont comparables, et dans certains cas supérieures, aux méthodes basées sur les algorithmes [1]. En revanche le temps de calcul pourrait poser un problème. Toutefois cette durée reste suffisamment faible pour ne pas constituer un réel handicap, et tout progrès informatique dans les temps de calcul serait un avantage immédiatement récupérable pour la méthode.

IV. L'intégrale de Boole [2]

A. Méthodologie générale

Nous avons testé comparativement, pour le calcul des intégrales, la méthode de SIMPSON et la méthode de BOOLE. Cette dernière, moins connue que la méthode de SIMPSON, s'étant révélée plus rapide, c'est celle que nous avons retenue, et que nous décrivons sommairement ci-après.

Soit $$I = \int_a^b f(x)\,dx$$

La méthode de BOOLE consiste à diviser l'intervalle d'intégration en n intervalles égaux, n étant un multiple de 4, et à lisser la courbe f(x) par une série d'arcs du 4^e ordre passant par les points de coordonnées x_0, $f(x_0)$; ... ; x_4, $f(x_4)$ pour le premier arc, x_4, $f(x_4)$; ... ; x_8, $f(x_8)$ pour le second etc. avec $x_0 = a$ et $x_n = b$.

La somme de BOOLE s'écrit :

$$S_n = \frac{1}{45}\frac{b-a}{n}\,[14\,C_0 + 64\,C_1(n) + 24\,C_2(n) + 28\,C_3(n)]$$

avec

$C_0 = f(a) + f(b)$

$C_1(n) = \sum f(a + i\,h)$ i impair, $h = \dfrac{b-a}{n}$

$C_2(n) = \sum f(a + i\,h)$ i pair, non multiple de 4 ($i \neq 0$ et $i \neq n$)

$C_3(n) = \sum f(a + i\,h)$ i multiple de 4 ($i \neq 0$ et $i \neq n$)

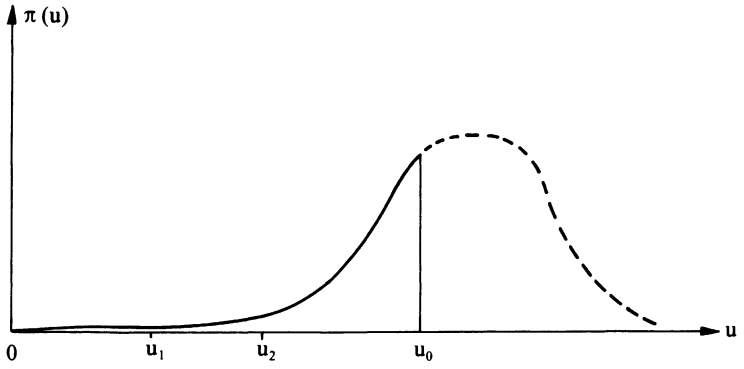
L'intégrale de BOOLE est en fait une moyenne pondérée. Elle peut s'écrire :

$$S_n = (b-a)\frac{\sum_{i=0}^{n} \alpha_i f(x_i)}{n} \quad \text{avec} \quad \sum_{i=0}^{n} \alpha_i = n$$

On utilisera cette méthode de façon itérative en calculant S_4, S_8, S_{16}, etc. les calculs étant facilités par le fait que les composantes de la somme S_n servent au calcul de celles de S_{2n}. La théorie de l'intégrale de BOOLE établit que les sommes S_n convergent vers la valeur exacte de l'intégrale. Un test approprié, dépendant de la précision demandée, sur l'expression $\left| \dfrac{S_{2n} - S_n}{S_{2n}} \right|$ permet d'arrêter les itérations.

B. Améliorations apportées à l'intégrale de Boole

Nous verrons plus loin que, par un choix judicieux entre F et $1 - F$, on fera en sorte que la fonction à intégrer soit toujours croissante dans l'intervalle d'intégration $[0, u_0]$ et soit de plus un infiniment petit au voisinage de $u = 0$. Par exemple $F_{BE}(p, q ; u_0) = \displaystyle\int_0^{u_0} \frac{u^{p-1}(1-u)^{q-1}}{B(p, q)}\,du$, u_0 étant inférieur à l'espérance $\dfrac{p}{p+q}$.



Le calcul à effectuer est $\int_0^{u_0} \pi(u)\, du$.

Dans une première phase, on limitera l'intégration à $\int_{u_1}^{u_0} \pi(u)\, du$, u_1 étant tel que la probabilité à gauche de u_1 soit négligeable. Par exemple $\int_0^{u_1} \pi(u)\, du < 10^{-30}$, si on décide de ne pas calculer les probabilités inférieures à 10^{-20}, mais d'exprimer dans ce cas le résultat par $F < 10^{-20}$ (10^{-30} est choisi pour réserver un nombre de chiffres significatifs suffisant à F, lorsque F est voisin de 10^{-20}, et supérieur à cette même valeur).

u_1 peut être commodément obtenu par une expression du type $u_1 = E(U) - 13\ ET(U)$ en tenant compte de l'assimilation à la loi normale (où E(U) désigne l'espérance de U, et ET(U) son écart-type).

Dans une deuxième phase, on limite l'intégration à $\int_{u_2}^{u_0} \pi(u)\, du$. Pour avoir u_2, on effectue un test au cours de la 3e itération de BOOLE pour que $(u_2 - u_1)\ \pi(u_2)$ qui est une majorante de $\int_{u_1}^{u_2} \pi(u)\, du$, soit négligeable devant $\int_{u_2}^{u_0} \pi(u)\, du$, compte-tenu de la précision demandée pour F.

Par l'introduction de ces valeurs u_1 et u_2, tous les calculs de densité sont efficaces à partir de la 3e itération de BOOLE, c'est-à-dire que leur contribution dans l'intégrale porte sur des chiffres significatifs dans le format de précision demandé. Sans ces améliorations, le nombre de calculs de densités serait majoré, parfois dans des proportions très importantes. On a pu constater par des essais que, dans le cas de la loi Bêta, le calcul de $\int_0^{u_0} \pi(u)\, du$ pour des valeurs élevées de p et q peut être de plusieurs minutes, alors que le calcul de $\int_{u_2}^{u_0} \pi(u)\, du$ est de quelques secondes.

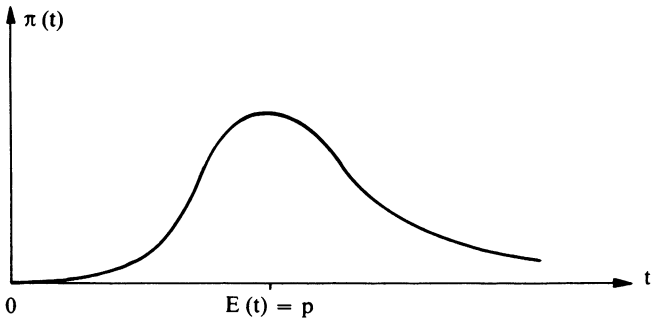
V. Précision des densités

A. Exemple de la distribution Gamma

Nous situerons nos analyses dans le contexte d'une calculatrice effectuant les calculs avec 12 chiffres en format virgule flottante (FVF). On distinguera l'écriture décimale habituelle (format décimal : FD) et l'écriture en FVF qu'on exprimera par :

$$x = x_0\ 10^{nx}, \quad \text{avec} \quad 0{,}1 \leqslant x_0 < 1$$

Pour étudier le problème général de la précision de la densité de probabilité, on s'appuiera sur l'exemple de la densité Gamma : $\pi_{GA}(p\ ;\ t) = \dfrac{e^{-t}\ t^{p-1}}{\Gamma(p)}$ (formule dans laquelle p est un nombre positif quelconque).



La densité est évaluée par son logarithme :

$$\text{Log}\ \pi_{GA}(p\ ;\ t) = -\ \log\ \Gamma(p) - t + (p - 1)\ \text{Log}\ t$$

Log Γ(p) est calculé par la formule de STIRLING [4] :

$$\text{Log}\ \Gamma(p) = \left(p - \frac{1}{2}\right)\ \text{Log}\,(p - 1) - (p - 1) + \frac{1}{2}\ \text{Log}\ 2\pi + \varepsilon p$$

$$\text{avec}\ \varepsilon p = \frac{B_1}{1.2(p - 1)} - \frac{B_2}{3.4(p - 1)^3} + \ldots +$$

$$(-\ 1)^{p-1}\ \frac{B_p}{(2p - 1)\ 2p\,.\,(p - 1)^{2p-1}} + \ldots$$

B_1, B_2, B_p sont les nombres de BERNOULLI.

Cette expression est le développement asymptotique, en série d'EULER MAC-LAURIN, de Log Γ(p) [4]. Bien que cette série soit divergente, elle donne, par un tronquage approprié, des résultats très précis. En effet, en utilisant la propriété des séries d'EULER MAC-LAURIN selon laquelle l'erreur due au tronquage est inférieure au double du premier terme négligé, on établit que, pour $p \geqslant 10$, et en tronquant au terme en B_5, l'erreur sur Log Γ(p) en FD est inférieure à 10^{-12}. Pour les valeurs de p inférieures à 10 on écrira

$$\Gamma(p) = \frac{\Gamma(p+i)}{p(p+1)\dots(p+i-1)} \quad \text{avec} \quad p + i \geqslant 10.$$

L'expression donnant Log $\pi_{GA}(p\,;t)$ peut être écrite :

$$\text{Log } \pi_{GA}(p\,;t) = A(p) - t + (p-1) \text{ Log } t, (^1)$$

avec :

$$A(p) = -\left(p - \frac{1}{2}\right) \text{Log } (p-1) + (p-1) - \frac{1}{2} \text{Log } 2\pi - \varepsilon p$$

(εp étant tronqué au terme en B_5).

On observe que cette densité est donnée par son logarithme. Il en sera de même pour la densité Bêta, donc pour toutes les distributions étudiées. Outre la commodité de l'expression analytique, le calcul de la densité par le logarithme permet, pour les densités très petites, de supprimer le risque de dépassement de capacité (UNDERFLOW), qui est important si on calcule directement les densités.

B. L'erreur de cadrage

Sur l'expression de Log π (2) on constate immédiatement le risque d'une importante perte de précision par erreur de cadrage. En effet, Log $\pi_{GA}(p\,;t)$ est compris entre -69 (on a vu qu'on assimile à zéro les probabilités inférieures à 10^{-30}, et Log $(10^{-30}) = -69$), et -2 ($\pi_{GA}(p\,;p-1)$, qui correspond au mode, décroît avec p et vaut 0,13 pour $p = 10$, et Log $(0{,}13) = -2$). Or le plus grand terme de Log π qui est $(p - 1/2)$ Log $(p-1)$ peut être très grand. Par exemple pour $p = 100\,000$, $(p - 1/2)$ Log $(p-1)$ est supérieur à 10^6 et sa partie entière a donc 7 chiffres. La perte de précision, par erreur de cadrage, est alors de 5 chiffres significatifs en FD, et Log π n'a que 5 décimales exactes.

L'erreur de cadrage sera souvent rencontrée. Pour la pallier il faudra chercher à améliorer l'écriture analytique de la densité.

C. Expression probabiliste des erreurs

En reprenant l'exemple ci-dessus, où Log π a 5 décimales exactes, et en désignant l'écart-type de l'erreur sur Log π par ET (Log π), on a pu écrire :

ET (Log π) $\leqslant k\,10^{-5}\,\sigma_0$, σ_0 étant l'écart-type de la VA de densité uniforme sur le segment $-0{,}5 + 0{,}5$ (soit 0,289), VA qui sert toujours de point de départ pour le calcul des erreurs. Log π, combinaison linéaire de VA, suit sensiblement une loi normale.

---

(1) Dans la programmation que nous retiendrons pour la distribution Gamma des améliorations seront apportées à l'écriture de Log $\pi_{GA}(p\,;t)$. Pour l'instant nous donnons cette formule à titre d'exemple.

(2) Pour ne pas alourdir les notations, on écrira π au lieu de $\pi(u)$ pour la densité. Ce qui ne devrait pas entraîner de confusion avec $\pi = 3{,}14$ qui, du reste, n'intervient plus par la suite.

k peut être évalué par des considérations probabilistes. Par exemple :

$ET(\text{Log}\,\pi) \leqslant 0{,}52\ 10^{-5}\,\sigma_0$ signifie que l'intervalle $[-\ 0{,}5\ 10^{-5}\ ;\ 0{,}5\ 10^{-5}]$ a une longueur supérieure à 6,6 $ET(\text{Log}\,\pi)$ et que, l'erreur sur $\text{Log}\,\pi$ étant normale, la probabilité pour que la valeur calculée de $\text{Log}\,\pi$ ait 5 décimales exactes est supérieure à 0,999, probabilité que nous assimilerons à la certitude.

Plus généralement, la relation :

$ET(\text{Log}\,\pi) \leqslant 0{,}52\ 10^{-n}\,\sigma_0$ implique que $\text{Log}\,\pi$ a n décimales exactes.

Tous nos calculs d'erreur ont été largement favorisés par des écritures de ce type, aussi bien pour les densités que pour l'intégrale de BOOLE.

Parallèlement, il est intéressant de noter qu'une inégalité sur l'écart-type de $\text{Log}\,\pi$ est équivalente à une inégalité sur l'écart-type relatif de π : $\frac{ET(\pi)}{\pi}$. En effet la relation : $\frac{d\pi}{\pi} = d\ \text{Log}\ \pi$ implique $\frac{ET(\pi)}{\pi} = ET(\text{Log}\ \pi)$. (Dans cette formule, on donne à la densité π au dénominateur du premier membre, le caractère d'une constante, compte-tenu de sa faible variabilité. On assimile ainsi π à $E(\pi)$).

VI. Précision des fonctions de répartition

On a vu que l'intégrale de BOOLE peut se mettre sous la forme

$$F = \left(\frac{u_0 - u_2}{n}\right) \sum_{i=0}^{n} \alpha_i \pi_i \text{ avec } \sum_{i=0}^{n} \alpha_i = n$$

(On a posé : $\pi_i = \pi(u_i)$).

A. Effet de l'erreur systématique de π

Cette erreur concerne les éléments de l'écriture de $\text{Log}\,\pi$ ne contenant pas la variable d'intégration. On exprimera cette erreur par son écart-type sur le logarithme, et le point de départ des calculs d'erreur sera toujours une inégalité préalablement établie de la forme :

$$ET_s(\text{Log}\,\pi) \leqslant \lambda\ 10^{-s}\,\sigma_0$$

Par différenciation de la formule donnant F on obtient :

$$dF = \left(\frac{u_0 - u_2}{n}\right) \Sigma\, \alpha_i\ d\pi_i \text{ avec } d\pi_i = \pi_i d\ \text{Log}\ \pi_i$$

D'où $dF = \left(\frac{u_0 - u_2}{n}\right) \Sigma\, \alpha_i \pi_i\ d\ \text{Log}\ \pi_i$. L'erreur étant systématique, $d\ \text{Log}\ \pi_i$ est indépendant de π_i et

$$ET_s(F) = \left(\frac{u_0 - u_2}{n}\right) (\Sigma\, \alpha_i \pi_i)\ ET_s(\text{Log}\ \pi)$$

De la relation $ET_s(\text{Log}\ \pi) \leqslant \lambda\ 10^{-s}\,\sigma_0$, et par division par

$$F = \left(\frac{u_0 - u_2}{n}\right) \Sigma\, \alpha_i \pi_i,$$

on obtient :

$$\frac{ET_s(F)}{F} \leqslant \lambda\, 10^{-s}\, \sigma_0$$

Exprimons F en FVF : $F = F_0\, 10^{nF}$

$$\frac{dF}{F} = \frac{dF_0}{F_0} \text{ soit } \frac{ET_s(F)}{F} = \frac{ET_s(F_0)}{F_0}$$

D'où : $\frac{ET_s(F_0)}{F_0} \leqslant \lambda\, 10^{-s}\, \sigma_0$ et, comme on a $0{,}1 \leqslant F_0 < 1$: $ET_s(F_0) \leqslant \lambda\, 10^{-s}\, \sigma_0$

On a ainsi établi :

$$ET_s(\text{Log}\,\pi) \leqslant \lambda\, 10^{-s}\, \sigma_0 \Rightarrow ET_s(F_0) \leqslant \lambda\, 10^{-s}\, \sigma_0$$

En d'autres termes, si Log π est entâché d'une erreur systématique limitant son exactitude à s décimales (FD), F est entâché d'une erreur limitant son exactitude à s chiffres (FVF).

B. Effet de l'erreur aléatoire de π

Cette erreur concerne les éléments de l'écriture de Log π contenant la variable d'intégration. On exprimera également cette erreur par son écart-type sur le logarithme, et le point de départ des calculs d'erreur sera toujours une inégalité préalablement établie de la forme :

$$ET_a(\text{Log}\,\pi) \leqslant \mu\, 10^{-a}\, \sigma_0$$

De la relation $F = \left(\frac{u_0 - u_2}{n}\right) \Sigma\, \alpha_i\, \pi_i$ on déduit, en raison de l'indépendance des erreurs aléatoires :

$$VAR_a(F) = \left(\frac{u_0 - u_2}{n}\right)^2 \Sigma\, \alpha_i^2\, VAR_a(\pi_i)$$

Tenant compte de la relation $ET_a(\pi_i) = \pi_i\, ET_a(\text{Log}\,\pi_i)$, on peut écrire :

$$VAR_a(F) = \left(\frac{u_0 - u_2}{n}\right)^2 \Sigma\, \alpha_i^2\, \pi_i^2\, VAR_a(\text{Log}\,\pi_i).$$

D'où :

$$VAR_a(F) \leqslant \left(\frac{u_0 - u_2}{n}\right)^2 (\Sigma\, \alpha_i^2\, \pi_i^2)\, \mu^2\, 10^{-2a}\, \sigma_0^2$$

$$ET_a(F) \leqslant \left(\frac{u_0 - u_2}{n}\right) \sqrt{\Sigma\, \alpha_i^2\, \pi_i^2}\, \mu\, 10^{-a}\, \sigma_0$$

En divisant par $F = \left(\frac{u_0 - u_2}{n}\right) \Sigma\, \alpha_i\, \pi_i$ on obtient :

$$\frac{ET_a(F)}{F} \leqslant \frac{\sqrt{\Sigma\, \alpha_i^2\, \pi_i^2}}{\Sigma\, \alpha_i\, \pi_i}\, \mu\, 10^{-a}\, s_0$$

En écrivant F en virgule flottante $F = F_0\, 10^{nF}$, et en posant $r = \frac{\sqrt{\Sigma\, \alpha_i^2\, \pi_i^2}}{\Sigma\, \alpha_i\, \pi_i}$ on obtient :

$$\frac{ET_a(F_0)}{F_0} \leqslant r\,\mu\,10^{-a}\,\sigma_0$$

De l'inégalité $\sqrt{\Sigma x_i 2} < \Sigma x_i$ ($x_i > 0$) on déduit $r < 1$. Au surplus des tests nous ont permis d'établir que le rapport r, dans les cas usuels utiles, est $< 1/5$. D'où le résultat :

$$ET_a(\text{Log}\,\pi) \leqslant \mu\,10^{-a}\,\sigma_0 \Rightarrow ET_a(F_0) \leqslant \frac{1}{5}\,\mu\,10^{-a}\,\sigma_0$$

Il y a compensation des erreurs aléatoires, qui se traduit par une amélioration de la précision, mais le gain est inférieur à un chiffre significatif. Au total, si Log π est entâché d'une erreur aléatoire limitant son exactitude à a décimales (FD), F est entâché d'une erreur limitant son exactitude à a chiffres (FVF).

C. Applications des calculs d'erreur

Il y a deux applications essentielles :

— D'une part ces calculs permettent d'établir la précision de F, caractéristique principale du logiciel. Bien entendu la précision a été vérifiée auprès des tables déjà existantes (Tables de PEARSON de la fonction Gamma incomplète et de la fonction Bêta incomplète), mais les calculs d'erreur sont la vraie validation.

— D'autre part ils sont à la base de la recherche de l'optimisation de l'écriture des densités : On s'efforcera d'abord de minimiser l'erreur de cadrage, principale source d'erreur, au besoin par un changement de la variable d'intégration, puis de doser autant que possible erreur systématique et erreur aléatoire, de telle sorte que les effets conjugués soient minimum, tout en conservant une écriture simple, car une écriture compliquée augmente le temps de calcul. Le gain ainsi obtenu par cette démarche, pour la précision de F, apparaîtra de façon plus nette dans l'étude de la programmation des lois Gamma et Bêta.

VII. Programmation de la loi Gamma

Nous ne donnons ici que les principes essentiels des calculs.

A. Cas $x \leqslant p - 1$

On calcule F après avoir effectué le changement de variable $u = \frac{t}{p-1}$ sur l'intégrale $F_{GA}(p\,;x) = \int_0^x \frac{e^{-t}\,t^{p-1}}{\Gamma(p)}\,dt$, formule dans laquelle le paramètre p est un nombre positif quelconque (toutefois nous limitons ses valeurs à $0{,}3 \leqslant p \leqslant 1\,000\,000$).

L'intégrale s'écrit alors :

$$F = \int_{u_2}^{u_0} f(u)\,du \text{ avec } f(u) = \frac{(p-1)^p}{\Gamma(p)}\,e^{-(p-1)u}\,u^{p-1} \text{ et } u_0 = \frac{x}{p-1}$$

Par le développement de STIRLING de Log (Γ (p)), on aboutit pour f (u) à l'expression

$$\text{Log } f(u) = A(p) - (p-1)(u - 1 - \text{Log } u)$$

A (p) étant une valeur comprise entre 0,1 et 6, et calculable avec précision, il en résulte que cette écriture est particulièrement intéressante, car elle supprime l'erreur de cadrage. L'erreur systématique sur A (p) est négligeable. Celle sur le terme $(p-1)(u-1-\text{Log } u)$ a le caractère d'une erreur aléatoire.

Nous avons pu, à partir de cette formule, avoir l'assurance d'obtenir pour F, et pour $0{,}3 \leqslant p \leqslant 1\,000\,000$ une précision de 7 chiffres en FVF. En réalité, pour $p < 1\,000$ le nombre de chiffres exacts est de 9 mais, pour obtenir 9, il faut consentir une durée de calcul plus longue. Nous avons préféré en rester au standard de 7 chiffres exacts en FVF.

On a pu vérifier que l'écriture de la densité par la formule $\text{Log } f(u) = A(p) - (p-1)(u-1-\text{Log } u)$, plutôt que par la formule brute initiale, fait gagner 3 chiffres dans la précision de F.

B. Cas $x > p - 1$

Nous calculons $F' = 1 - F = \int_x^\infty \frac{e^{-t} t^{p-1}}{\Gamma(p)}\, dt$. En effectuant le changement de variable $u = \frac{p-1}{t}$, on aboutit à $F' = \int_{u_2}^{u_0} g(u)\, du$, avec $u_0 = \frac{p-1}{x}$.

Log g (u) se développe selon une formule analogue à celle relative à Log f (u). Log g (u) jouit du point de vue de la distribution de l'erreur systématique et de l'erreur aléatoire des mêmes avantages que Log f (u) précédemment étudié, et ses performances, en terme de précision sont les mêmes.

En distinguant les cas $x \leqslant p - 1$, $x > p - 1$, on est donc en mesure d'exprimer à la fois F et $1 - F$ avec la précision demandée (en format virgule flottante).

C. Cas particulier des petites valeurs de p

Pour $0{,}3 \leqslant p < 10$, deux difficultés de programmation s'ajoutent. D'une part Log $\Gamma(p)$ n'est plus calculable par la formule de STIRLING, d'autre part la faiblesse de l'exposant de t^{p-1} dans $\pi(t)$ pour $p < 2$ entraîne une convergence lente pour l'intégrale de BOOLE, voire une divergence.

Ces difficultés ont été levées en utilisant la relation :

$$\Gamma(p) = \frac{\Gamma(p+i)}{p(p+1)\ldots(p+i-1)} \text{ avec } p + i > 10$$

et le développement bien connu de $F_{GA}(p\,;x)$, obtenu par intégrations successives par parties

$$\int_0^x \frac{e^{-t}\,t^{p-1}}{\Gamma(p)}\,dt = \frac{e^{-x}\,x^p}{p\Gamma(p)} + \frac{e^{-x}\,x^{p+1}}{p(p+1)\,\Gamma(p)} + \ldots +$$

$$+ \ldots + \frac{e^{-x}\,x^{p+i}}{p(p+1)\ldots(p+i)\,\Gamma(p)} + \int_0^x \frac{e^{-t}\,t^{p+i}}{p(p+1)\ldots(p+i)\,\Gamma(p)}\,dt$$

D. Bilan de la programmation

Tous ces calculs ont pu être intégrés dans un programme unique, commun avec celui de la loi Bêta, avec blocs, sous-routines etc. qui permet donc le calcul de F et 1 − F avec 7 chiffres exacts en FVF pour $0{,}3 \leqslant p \leqslant 1\,000\,000$.

VIII. Programmation de la loi Bêta

Comme pour la loi Gamma nous avons calculé, selon le cas, F ou 1 − F.

A. Choix du calcul de F ou 1 − F

On utilise la relation :

$$F_{BE}(p, q ; x) = 1 - F_{BE}(q, p ; (1 - x))$$

Le mode opératoire est donc simple :

pour $x \leqslant \frac{p}{p+q}$ $\left(E(X) = \frac{p}{p+q}\right)$ on calculera directement $F_{BE}(p, q ; x)$

pour $x > \frac{p}{p+q}$ on calculera $F_{BE}(q, p ; (1 - x)) = 1 - F_{BE}(p, q ; x)$.

B. Expression de la densité

$$F_{BE}(p, q ; u_0) = \int_{u_2}^{u_0} \pi(u)\,du$$

avec $$\pi(u) = \frac{u^{p-1}(1-u)^{q-1}}{B(p, q)} = \frac{\Gamma(p+q)}{\Gamma(p)\,\Gamma(q)}\,u^{p-1}(1-u)^{q-1}$$

Les paramètres p et q sont des nombres positifs quelconques. Toutefois nous limiterons leurs valeurs à $0{,}3 \leqslant \frac{p}{q} \leqslant 1\,000\,000$.

On calculera $\pi(u)$ par son logarithme. On pourrait calculer $\text{Log}\left(\frac{1}{B(p, q)}\right)$ par la formule de STIRLING appliquée 3 fois, de façon indépendante, à Log Γ (p), Log Γ (q), Log Γ (p + q). Il est plus intéressant de simplifier l'écriture globale $\text{Log}\left(\frac{1}{B(p, q)}\right) = -\text{Log}\,\Gamma(p) - \text{Log}\,\Gamma(q) + \text{Log}\,\Gamma(p+q)$ exprimée par 3 développements de STIRLING. On aboutit ainsi à une expression de la forme :

$$\text{Log}\left(\frac{1}{B(p,q)}\right) = H(p) + (p-1)\,\text{Log}\left(\frac{p+q-1}{p-1}\right) + (q-1)\,\text{Log}\left(\frac{p+q-1}{q-1}\right)$$

H (p) a une valeur de quelques unités, et sa précision ne pose pas de problème. Les deux autres expressions sont moins précises. Si on leur applique les calculs d'erreur il apparaît que, pour p et q de l'ordre de 1 000 000, $\text{Log}\left(\frac{1}{B(p,q)}\right)$ a 5 décimales exactes seulement.

Au total nous écrirons la densité sous la forme :

$$\text{Log}\,\pi(u) = H(p) + (p-1)\,\text{Log}\left(\frac{p+q-1}{p-1}\right) + (q-1)\,\text{Log}\left(\frac{p+q-1}{q-1}\right) + (p-1)\,\text{Log}\,u + (q-1)\,\text{Log}(1-u)$$

Nous avons tenté une combinaison des 2e et 4e termes d'une part et des 3e et 5e termes d'autre part, mais elle s'est heurtée à des difficultés de programmation, et s'est révélée moins efficace qu'il y paraît sur le plan de la précision.

C'est donc la formule ci-dessus que nous avons utilisée en programmation; les 3 premiers termes engendrant l'erreur systématique, les 2 derniers termes l'erreur aléatoire. On a pu établir que cette écriture permet, par rapport à l'écriture brute de Log π (u), d'améliorer la précision de F de 2 chiffres.

C. Cas des petites valeurs de p et/ou de q

La procédure suivie a été analogue à celle de la loi GAMMA : d'une part nous avons programmé la formule :

$$\Gamma(p) = \frac{\Gamma(p+i)}{p(p+1)\ldots(p+i-1)} \quad (p+i \geqslant 10)$$

et d'autre part nous avons utilisé pour p (resp q) $\leqslant 2$ un développement de la forme :

$F_{BE}(p,q;x) = a_1(p,q,x) + a_2(p,q,x) + \ldots + F_{BE}(p+i, q-i; x)$ en faisant en sorte que p + i soit supérieur à 2.

D. Bilan de la programmation

Le programme ainsi établi, commun avec celui de la loi Gamma, permet le calcul de F et 1 − F pour $0{,}3 \leqslant \frac{p}{q} \leqslant 1\,000\,000$.

Toutefois la précision est inférieure à celle obtenue pour Gamma, à savoir :

$p + q \leqslant 50\,000$	7 chiffres exacts en FVF
$50\,000 < p + q \leqslant 200\,000$	6 chiffres exacts en FVF
$200\,000 < p + q \leqslant 2\,000\,000$	5 chiffres exacts en FVF

IX. Conclusion

Nous nous bornerons à signaler, en conclusion, les possibilités d'extension de cette méthode à d'autres lois de probabilités de la variable continue. La plupart de ces lois peuvent à priori faire l'objet d'un traitement analogue, mais la programmation reste délicate, car les difficultés relatives au dépassement de capacité, à l'écriture des densités, à l'assurance de couvrir un large domaine de variation des paramètres, au contrôle de la précision etc., sont nombreuses, et doivent être résolues de façon spécifique pour chaque loi.

Nous rappelons également qu'avec des langages informatiques plus performants actuellement disponibles, la précision de 7 chiffres en format virgule flottante serait accessible partout, et le temps de calcul serait sensiblement réduit.

Bibliographie

[1] M. ABRAMOWITZ, I.A. STEGUN. — Handbook of mathematical fonctions, Dover Publication Inc., New-York 1972.

[2] R. BRARD. — Cours de mathématiques appliquées — Ecole Polytechnique.

[3] B. LAUMON, J.L. MARTIN, D. BALLAND. — Un algorithme de calcul de la probabilité qu'une variable de Fisher-Snedecor, à premier degré de liberté impair, excède une valeur donnée, *Revue de Statistique Appliquée*, 1985, vol. XXXIII, n° 4.

[4] F. SCHEID. — *Analyse numérique*, Série SCHAUM Mc Graw-Hill.