# études

## de l'utilité d'un tri préalable ou somme de deux variables normales "tronquées"

*par E. Bernay*
*Lycée Chevrollier, Angers*

Nous trouvons dans un ouvrage de Mathématiques appliquées (réf. Probabilités - initiation à la recherche opérationnelle de C. GOUJET et C. NICOLAS aux Editions Masson) l'exercice suivant :

"Une machine à empaqueter des bonbons prépare des sachets de 25 g ; leur poids réel est en fait une variable aléatoire de moyenne 25 et d'écart type 2 (loi normale).

Un distributeur commande des paquets de 50 g et refuse systématiquement ceux dont le poids est inférieur à 46 g. Pour préparer ces paquets de 50 g sans désorganiser le mode actuel d'empaquetage, deux solutions sont envisagées :

a) réunir deux sachets de 25 g. Quel est alors le pourcentage de paquets qui seront renvoyés par le distributeur ?

b) répartir d'abord les sachets de 25 g en deux classes : ceux de plus de 25 g et ceux de moins de 25 g. Puis on prend un sachet dans la première classe et on le réunit avec un sachet de la deuxième classe. Quelle sera alors la proportion de paquets refusés par le distributeur ?"

La question a) est classique et se résoud sans difficultés car la somme de deux aléas normaux indépendants est un aléa normal (cf. annexe I). Par contre, la question b) amène à se poser le problème suivant : en additionnant deux aléas normaux "tronqués" définis par exemple par leurs densités :

• pour le 1er aléa $T_1$ :

$$\begin{cases} f_1(x) = 2n(x) = \dfrac{2}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}} & \text{pour } x<0 \\ f_1(x) = 0 & \text{pour } x\geqslant 0 \end{cases}$$

• pour le 2e aléa $T_2$ :

$$\begin{cases} f_2(x) = 0 & \text{pour } x<0 \\ f_2(x) = 2n(x) = \dfrac{2}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}} & \text{pour } x\geqslant 0 \end{cases}$$

obtient-on un aléa normal ?

On verra que la réponse est négative (annexe II) mais que l'approximation normale fournit un résultat numérique "proche" de la réalité (annexe I).

Dans *l'annexe I*, nous répondons d'abord à la question a) puis à la question b) en faisant l'hypothèse (fausse en réalité) que la somme des deux aléas "normaux tronqués" est normale : il suffira de calculer le nouvel écart-type qui, étant plus petit que le précédent (le facteur est $\sqrt{2-\frac{4}{\pi}} = 0{,}8525$), explique la proportion moindre de paquets refusés et justifie l'utilité d'un tri préalable.

Dans *l'annexe II*, nous répondons de façon précise à la question b) en déterminant successivement la densité (par un produit de convolution) puis la fonction de répartition (calcul d'une intégrale double "généralisée" de la somme des deux aléas "normaux tronqués". Nous retrouvons également l'écart-type et son facteur $\sqrt{2-\frac{4}{\pi}}$. La réponse numérique se fait par le calcul approché d'une intégrale généralisée (annexe III).

Dans *l'annexe III*, nous comparons les résultats numériques des questions a) et b) (approximation normale de l'annexe I, calcul direct de l'annexe II) en prenant successivement comme limite inférieure de poids acceptés par le distributeur (poids en grammes) : 44, 45, *46*, 47, 48, 49, 50.
(énoncé)

Sur les trois tronquées on pourra consulter :

L. MAILHOT, **Etude des lois de probabilités réelles tronquées à droite et applications statistiques.** Thèse de 3e cycle, Clermont II, 1985.

## ANNEXE I

**Notation :**

T désignera une variable gaussienne centrée réduite (T suit la loi normale $\mathcal{N}(0;1)$) dont la fonction de répartition (tabulée) sera notée :

$N : x \to N(x) = \text{Prob}(T \leqslant x)$ et la densité $n : x \to n(x) = \dfrac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}$.

**Réponse a) :**

$X_1$ désignant la variable aléatoire du poids du premier sachet, $X_1$ suit la loi normale $\mathcal{N}(25;2)$. Il en est de même pour $X_2$, variable aléatoire du poids du deuxième sachet. $X_1$ et $X_2$ étant indépendants, $X_1+X_2$ suit la loi normale $\mathcal{N}(50;2\sqrt{2})$, donc :

$$\text{Prob}\,(X_1+X_2\leqslant 46) = P\left(T\leqslant \frac{46-50}{2\sqrt{2}} = -\sqrt{2}\right) = N(-1{,}414) = 0{,}079$$

soit environ *8 % de paquets refusés.*

**Réponse b) :**

$X_1$ désignant la variable aléatoire du poids du sachet de la première classe (plus de 25 g) et X la variable aléatoire du poids d'un sachet initial (X suit la loi $\mathcal{N}(25;2)$), on a :

pour $x\leqslant 25$, $P(X_1<x)=0$

pour $x>25$, $P(X_1<x)=P(X<x/X>25)=\dfrac{P(25<X<x)}{P(X>25)}$

$$=\frac{F(x)-F(25)}{1-F(25)}\quad (F : \text{fonction de répartition de X})$$

$$=2F(x)-1 \qquad \text{car } F(25)=0{,}5$$

d'où la densité $f_1$ de $X_1$ :

$$\begin{cases}\text{pour } x\leqslant 25,\ f_1(x)=0\\ \text{pour } x>25,\ f_1(x)=2f(x)\end{cases}$$ où f est la densité de X

un calcul analogue donne la densité $f_2$ de $X_2$ :

$$\begin{cases}\text{pour } x\leqslant 25,\ f_2(x)=2f(x)\\ \text{pour } x>25,\ f_2(x)=0\end{cases}$$

en utilisant les résultats classiques :

$$\int_0^{+\infty} e^{-\frac{x^2}{2}}dx=\sqrt{\frac{\pi}{2}},\quad \int_0^{+\infty} xe^{-\frac{x^2}{2}}dx=1 \quad \text{et} \quad \int_0^{+\infty} x^2e^{-\frac{x^2}{2}}dx=\sqrt{\frac{\pi}{2}}$$

on détermine successivement les paramètres de ces deux lois, soit

$$E(X_1)=m+\frac{2\sigma}{\sqrt{2\pi}}\ (\text{ici } m=25 \text{ et } \sigma=2),\quad E(X_2)=m-\frac{2\sigma}{\sqrt{2\pi}},$$

$$V(X_1)=V(X_2)=\sigma^2\left(1-\frac{2}{\pi}\right).$$

$X_1$ et $X_2$ étant indépendantes, $E(X_1+X_2)=E(X_1)+E(X_2)=2m\ (=50)$

et $V(X_1+X_2)=V(X_1)+V(X_2)=\sigma^2\left(2-\dfrac{4}{\pi}\right)$

ou $\sigma_{X_1+X_2}=\sigma\sqrt{2-\dfrac{4}{\pi}}\ (=1{,}705)$.

En *admettant* que $X_1 + X_2$ suit une loi normale (faux en réalité)

$$\text{Prob}\,(X_1 + X_2 < 46) = P\left(T < \frac{46-50}{1{,}705}\right) = N(-2{,}346) \simeq 0{,}0095$$

soit *environ 1 % de paquets refusés.*

**Généralisation :**

On vient d'établir que si T suit une loi normale et si $X_1$, $X_2$ sont obtenus en tronquant T à la moyenne (qui est ici la médiane)

$$V(X_1 + X_2) < 2V(T).$$

Or, ce résultat est encore vrai si T a une fonction de répartition F quelconque, en effet : soit a un réel et $\alpha$ un nombre compris entre F(a) et F(a + 0) ($0 < \alpha < 1$).

Les deux répartitions tronquées associées à a et $\alpha$ sont :

$$G(x) = \begin{cases} \dfrac{F(x)}{\alpha} & x \leqslant a \\ 1 & x > a \end{cases}$$

$$D(x) = \begin{cases} \dfrac{F(x)}{1-\alpha} & x > a \\ 0 & x \leqslant a \end{cases}$$

Soient $X_1$ et $X_2$ deux variables indépendantes de lois respectives G et D. On a :

$$EX_1 = \frac{1}{\alpha} \int_{]-\infty, a[} x\, dF(x) + \frac{\alpha - F(a)}{\alpha}\, a$$

$$EX_2 = \frac{1}{1-\alpha} \int_{]a, +\infty[} x\, dF(x) + \frac{F(a+0) - \alpha}{1-\alpha}\, a$$

de sorte que $X_1 + X_2 = 2 \int_{-\infty}^{+\infty} x\, dF(x) = 2ET$ si $\alpha = 1 - \alpha = \frac{1}{2}$

c'est-à-dire si $a$ est une *médiane* de T, alors

$$EX_1^2 = 2 \int_{]-\infty, a[} x^2\, dF(x) + (1 - 2F(a))\, a^2$$

$$EX_2^2 = 2 \int_{]a, +\infty[} x^2\, dF(x) + (2F(a+0) - 1)\, a^2$$

et $\text{var}\,(X_1 + X_2) = E_1^2 + EX_2^2 - (EX_1)^2 - (EX_2)^2 = 2ET^2 - (EX_1)^2 - (EX_2)^2$

d'où $2\,\text{var}\,T - \text{var}\,(X_1 + X_2) = (EX_1)^2 + (EX_2)^2 - 2(ET)^2$

$$= \frac{1}{2}(EX_2 - EX_1)^2 > 0$$

sauf si T est concentrée en $a$.

# ANNEXE II

## 1. Densité de la somme de deux aléas normaux "tronqués" centrés indépendants

Soit $T_1$ l'aléa normal "tronqué" centré de densité $f_1$ définie par :

$$\begin{cases} f_1(t) = \dfrac{2}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} & \text{si et seulement si } t<0 \\ f_1(t) = 0 & \text{si et seulement si } t\geqslant 0 \end{cases}$$

et $T_2$ l'aléa normal "tronqué" centré de densité $f_2$ définie par :

$$\begin{cases} f_2(t) = 0 & \text{si et seulement si } t<0 \\ f_2(t) = \dfrac{2}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} & \text{si et seulement si } t\geqslant 0 \end{cases}$$

alors $T_1 + T_2$ a pour densité le produit de convolution de $f_1$ et de $f_2$ défini par :

$$(f_1 * f_2)(x) = f(x) = \int_{-\infty}^{+\infty} f_1(x-t)\, f_2(t)\, dt$$

Le résultat s'obtient sans difficultés :

$$f(x) = \frac{2}{\pi} e^{-\frac{x^2}{4}} \int_{\frac{|x|}{2}}^{+\infty} e^{-u^2}\, du$$

## 2. Graphe de la densité

— f est paire

— $f(0) = \dfrac{1}{\sqrt{\pi}}$

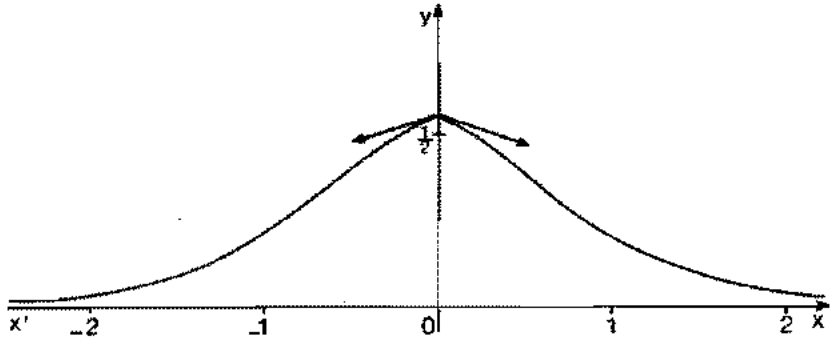
— f dérivable sur $\mathbb{R}^*$ et pour $x>0$ :

$$f'(x) = \frac{2}{\pi} e^{-\frac{x^2}{4}} \left[ -\frac{x}{2} \int_{\frac{x}{2}}^{+\infty} e^{-u^2}\, du - \frac{e^{-\frac{x^2}{2}}}{2} \right]$$

soit $f'(x)<0$ et $f'_d(0) = -\dfrac{1}{\pi}$

| $x$ | $0$ | | $+\infty$ |
|---|---|---|---|
| $f'(x)$ | $-\frac{1}{\pi}$ | — | |
| $f(x)$ | $\frac{1}{\sqrt{\pi}}$ | ↘ | $0$ |

| $x$ | $f(x$ |
|---|---|
| 0 | 0,564 |
| 0,5 | 0,384 |
| 1 | 0,21 |
| 1,5 | 0,093 |
| 2 | 0,033 |




*Remarque* : f n'est pas la densité d'une variable gaussienne puisque sa dérivée présente un saut en 0. Là encore ce résultat est général : notons comme ci-dessus F la fonction de répartition de T et $X_1$, $X_2$ les variables obtenues en tronquant en $a$ quelconque.

Supposons en outre F absolument continue et de densité f dérivable et à dérivée continue ; les densités de G et D sont :

$$g(x) = \frac{f(x)}{\alpha} 1_{]x \leqslant a]} \quad \text{et} \quad d(x) = \frac{f(x)}{1-\alpha} 1_{]x>a[}$$

et celle de $X_1 + X_2$ est :

$$\varphi(x) = \int_{-\infty}^{+\infty} g(t)\, h(x-t)\,dt = \int_{-\infty}^{+\infty} h(t)\, g(x-t)\,dt$$

$$= \begin{cases} \dfrac{1}{\alpha(1-\alpha)} \displaystyle\int_{-\infty}^{a} f(t)\, f(x-t)\,dt & \text{si } x \geqslant 2a \\ \dfrac{1}{\alpha(1-\alpha)} \displaystyle\int_{a}^{+\infty} f(t)\, f(x-t)\,dt & \text{si } x \leqslant 2a \end{cases}$$

de sorte que

$$\varphi'(x) = \begin{cases} \dfrac{1}{\alpha(1-\alpha)} \displaystyle\int_{-\infty}^{a} f(t)\, f'(x-t)\,dt & \text{si } x \geqslant 2a \\ \dfrac{1}{\alpha(1-\alpha)} \displaystyle\int_{a}^{+\infty} f(t)\, f'(x-t)\,dt & \text{si } x \leqslant 2a \end{cases}$$

et que $\lim\limits_{x \uparrow 2a} \varphi'(x) = \dfrac{1}{\alpha(1-\alpha)} \displaystyle\int_{a}^{+\infty} f(t)\, f'(2a-t)\,dt = \mu$

alors que $\lim\limits_{x \downarrow 2a} \varphi'(x) = \dfrac{1}{\alpha(1-\alpha)} \displaystyle\int_{-\infty}^{a} f(t)\, f'(2a-t)\,dt = \nu$

En intégrant par parties on obtient :

$$\mu = -\frac{f(t)\, f(2a-t)}{\alpha(1-\alpha)}\Bigg]_{a}^{+\infty} + \frac{1}{\alpha(1-\alpha)} \int_{a}^{+\infty} f'(t)\, f(2a-t)at$$

et en faisant le changement de variable $t = 2a - t'$

$$\mu = \nu + \frac{f^2(a)}{\alpha(1-\alpha)}$$

si donc $f(a) \neq 0$ $\varphi'$ a un saut au point d'abscisse $a$.

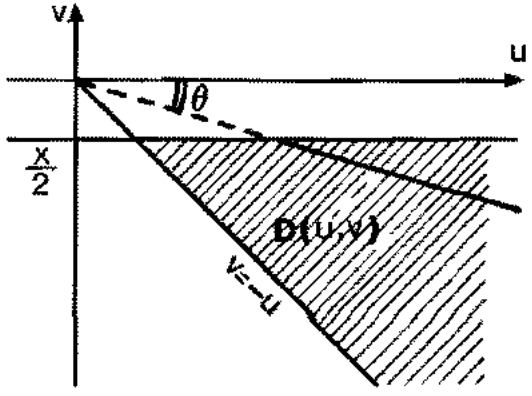
**3. Fonction de répartition**

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = \frac{2}{\pi} \int_{-\infty}^{x} e^{-\frac{t^2}{4}} \left( \int_{\frac{|t|}{2}}^{+\infty} e^{-u^2} du \right)$$

$$= \frac{4}{\pi} \int_{-\infty}^{\frac{x}{2}} e^{-v^2} \left( \int_{|v|}^{+\infty} e^{-u^2} du \right) dv \qquad \left[ v = \frac{t}{2} \right]$$

$$= \frac{4}{\pi} \iint_{D(u,v)} e^{-(u^2+v^2)}\, du\, dv \quad \text{où} \quad D(u,v) = \left\{ (u,v) \;/\; \begin{matrix} v \leqslant \frac{x}{2} \\ |v| \leqslant u \end{matrix} \right\}$$

représenté ci-dessous.



*pour* $x<0$ : en coordonnées polaires

$$F(x)=\frac{4}{\pi}\int_{-\frac{\pi}{4}}^{0}\left[\int_{\frac{x}{2\sin\theta}}^{+\infty}\rho\, e^{-\rho^2}\,d\rho\right]d\theta$$

$$=\frac{2}{\pi}\int_{0}^{\frac{\pi}{4}} e^{-\frac{x^2}{4\sin^2\theta}}\,d\theta$$

ou $\quad F(x)=\frac{1}{\pi}\int_{x}^{+\infty}\frac{e^{-\frac{x^2}{4}}\,u}{u\sqrt{u-1}}\,du \qquad \left(u=\frac{1}{\sin^2\theta}\right)$

et pour des raisons de symétrie: $F(0)=\frac{1}{2}$ et $F(x)=1-F(-x)$ pour $x>0$.

**4. Calcul de l'écart-type**

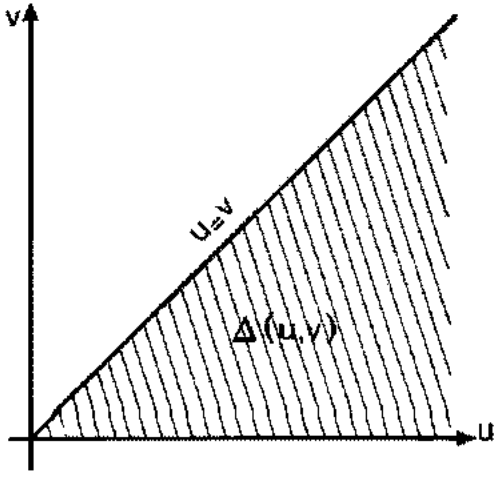
Soit $X_0=T_1+T_2$ alors évidemment $E(X_0)=0$ et

$$E(X_0^2)=\int_{\infty}^{+\infty}x^2 f(x)\,dx=\frac{4}{\pi}\int_{0}^{+\infty}x^2 e^{-\frac{x^2}{4}}\left[\int_{\frac{x}{2}}^{+\infty}e^{-u^2}\,du\right]dx$$

$$=\frac{32}{\pi\rho}\int_{0}^{+\infty}v^2 e^{-v^2}\left[\int_{v}^{+\infty}e^{-u^2}\,du\right]dv \quad (v=\frac{x}{2})$$

$$=\frac{32}{\pi}\iint_{\Delta(u,v)}v^2\, e^{-(u^2+v^2)}\,du\,dv$$

où $\quad \Delta(u,v)=\{(u,v)\ /\ 0\leqslant v\leqslant u\}$ ci-contre



en coordonnées polaires :

$$E(X_0^2) = \frac{32}{\pi} \int_0^{\frac{\pi}{4}} \sin^2\theta \left( \int_0^{+\infty} \rho^3 e^{-\rho^2}\, d\rho \right) d\theta$$

par parties : $\displaystyle\int_0^{+\infty} \rho^3 e^{-\rho^2}\, d\rho = \int_0^{+\infty} \rho\, e^{-\rho^2}\, d\rho = \frac{1}{2}$

et $\displaystyle\int_0^{\frac{\pi}{4}} \sin^2\theta\, d\theta = \int_0^{\frac{\pi}{4}} \left( \frac{1-\cos 2\theta}{2} \right) d\theta = \frac{\pi}{8} - \frac{1}{4}$

d'où $E(X_0^2) = 2 - \frac{4}{\pi}$ et on en déduit : $\sigma = \sqrt{2 - \frac{4}{\pi}}$

## ANNEXE III

**Résultats numériques**

1) Rappelons la fonction de répartition de la somme de deux aléas "tronqués" normaux, centrés et indépendants, calculée dans l'annexe II :

$$F(x) = P(X_0 \leqslant x) = \frac{1}{\pi} \int_2^{+\infty} \frac{e^{-\frac{x^2}{4}u}}{u\sqrt{u-1}}\, du \qquad \text{pour } x < 0$$

et son écart-type $\sigma_{X_0} = \sqrt{2 - \frac{4}{\pi}}$ .

Dans l'énoncé initial, en appelant X la somme des aléas normaux "tronqués" $X_1$ et $X_2$ on a vu que $E(X) = 50$ et $\sigma_X = 2\sqrt{2 - \frac{4}{\pi}}$ :

donc $X_0 = \frac{X - 50}{2}$ et $P(X \leqslant k) = P(X_0 \leqslant \frac{k-50}{2}) = F\,\frac{(k-50)}{2}$ .

Ainsi, si $k \in \{44, 45, 46, 47, 48, 49, 50\}$

$$x = \frac{k-50}{2} \in \{-3\ ;\ -2{,}5\ ;\ -2\ ;\ -1{,}5\ ;\ -1\ ;\ -0{,}5\ ;\ 0\} .$$

2) L'étude analytique de F pourrait donner lieu à des développements intéressants ; en ce qui concerne le calcul numérique de l'intégrale pour les valeurs de $x$ ci-dessus, j'ai "négligé" le "reste" :

$$\int_A^{+\infty} \frac{e^{-\frac{x^2}{4}u}}{u\sqrt{u-1}}\, du \leqslant \frac{4}{x^2 A\sqrt{A-1}}\, e^{-\frac{x^2}{4}A}$$

et utilisé la méthode de SIMPSON de la CASIO FX 180 P pour calculer :

$$\int_2^A \frac{e^{-\frac{x^2}{4}u}}{u\sqrt{u}-1}\,du$$

Le tableau ci-dessous donne successivement les valeurs de $k$ puis de $x$ (cf. ci-dessus) puis de A et de N (nombre de divisions de l'intervalle [2,A])

| $k$ | 44 | 45 | 46 | 47 | 48 | 49 |
|---|---|---|---|---|---|---|
| $x$ | $-3$ | $-2,5$ | $-2$ | $-1,5$ | $-1$ | $-0,5$ |
| A | 5 | 8 | 12 | 16 | 32 | 128 |
| N | 128 | 518 | 512 | 512 | 512 | 512 |

($x=0$ ne figure pas car $F(0)=\frac{1}{2}$)

3) Le tableau ci-dessous donne pour les valeurs de $k$ précédentes successivement :

— la réponse à la question a) (''pas de tri'')
— la réponse à la question b) (tri préalable)

1. par l'approximation normale (annexe I)
2. par le calcul direct de $F(x)$ (cf. ci-dessous)

| $k$ | 44 | 45 | 46 | 47 | 48 | 49 | 50 |
|---|---|---|---|---|---|---|---|
| sans tri | 0,017 | 0,0385 | 0,079 | 0,144 | 0,24 | 0,362 | 0,5 |
| avec tri normale | 0,00022 | 0,0017 | 0,0095 | 0,0392 | 0,12 | 0,279 | 0,5 |
| avec tri $F(x)$ | 0,00057 | 0,00297 | 0,012 | 0,0417 | 0,115 | 0,262 | 0,5 |

On constate bien l'utilité d'un tri préalable et le risque d'une approximation normale ''hâtive'' même ''proche'' de la réalité.