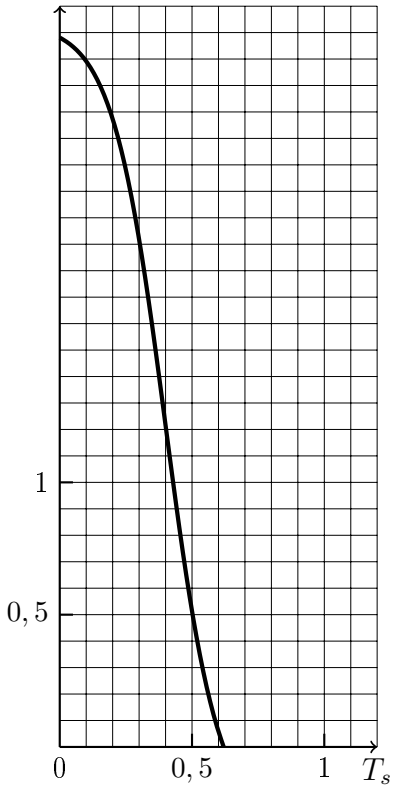
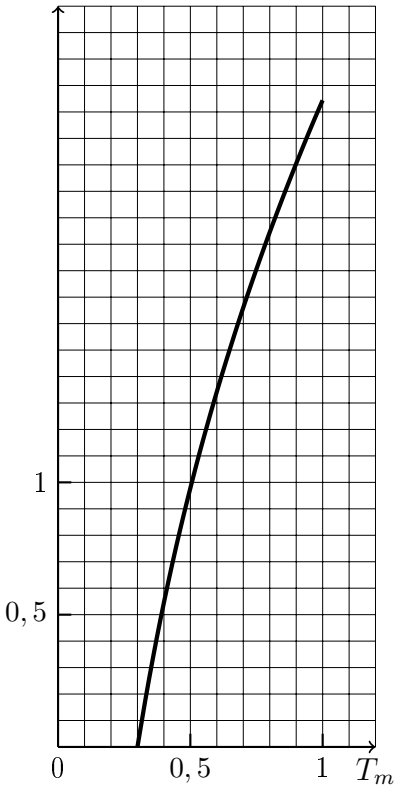
Corrigé du devoir

Exercice 1 : test médical (5 pts)

Afin de détecter une certaine maladie, on utilise un test médical dont le résultat est une valeur comprise entre 0 et 1.

Des expériences ont permis d'établir la répartition statistique des résultats du test chez les individus sains et chez les individus malades. On note T_s le résultat du test chez un individu sain et T_m celui d'un individu malade. Les graphes ci-dessous représentent les densités de T_s et T_m (un carreau$= 0,1 \times 0,1$).





1. Estimer grossièrement les espérances $\mathbf{E}(T_s)$ et $\mathbf{E}(T_m)$.

   Graphiquement, il semble évident que l'espérance de T_s est comprise entre $0,1$ et $0,3$. On peut l'estimer par $\mathbf{E}(T_s) \approx 0,2$. (La valeur exacte est $\mathbf{E}(T_s) \approx 0,211$.)
   On peut estimer que l'espérance de T_m est comprise entre $0,7$ et $0,8$. (La valeur exacte est $\mathbf{E}(T_s) \approx 0,746$.)

2. Proposer une valeur c telle que $\mathbf{P}(T_m > c) \approx 0,97$.

   En comptant les carreaux, on peut estimer que $\mathbf{P}(T_m < 0,4) \approx 0,03$. On en déduit que $c \approx 0,4$.

3. Estimer alors $\mathbf{P}(T_s > c)$.

   Sur le premier graphe, on compte le nombre de carreaux situés sous la courbe entre $0,4$ et $0,6$. On peut estimer qu'il y a 10-11 carreaux. Ainsi $\mathbf{P}(T_s > 0,4) \approx 0,11$.

   On utilise c comme valeur seuil : lorsque le test d'un individu fournit une valeur supérieure à c, il est déclaré positif, sinon il est déclaré négatif.

4. Sachant que cette maladie touche un individu sur cent, estimer les deux risques d'erreur : la probabilité d'être malade lorsque le test est négatif et la probabilité d'être sain lorsque le test est positif.

   Il faut modéliser le problème à l'aide d'un arbre. Notons $\ominus$ et $\oplus$ le fait que le test soit négatif ou positif. Alors

$$\mathbf{P}(\ominus) = \mathbf{P}(\ominus \text{ et } M) + \mathbf{P}(\ominus \text{ et } S) \approx 0,03 \cdot 0,01 + 0,89 \cdot 0,99 \approx 0,88.$$

$$\mathbf{P}(\oplus) = \mathbf{P}(\oplus \text{ et } M) + \mathbf{P}(\oplus \text{ et } S) \approx 0,97 \cdot 0,01 + 0,11 \cdot 0,99 \approx 0,12.$$

   Utilisons la formule de Bayes :

$$\mathbf{P}(M|\ominus) = P(\ominus|M)\frac{\mathbf{P}(M)}{\mathbf{P}(\ominus)} \approx 0,03 \cdot \frac{0,01}{0,88} \approx 0,0003.$$

$$\mathbf{P}(S|\oplus) = P(\oplus|S)\frac{\mathbf{P}(S)}{\mathbf{P}(\oplus)} \approx 0,11 \cdot \frac{0,99}{0,12} \approx 0,91.$$

   Conclusion : le test est très fiable lorsqu'il est négatif, mais en contrepartie, être testé positif ne signifie pas souvent qu'on est réellement malade.

Exercice 2 : simulation de la loi normale (15 pts)

Le but de cet exercice est d'étudier des méthodes de simulation numérique d'une variable aléatoire suivant la loi $\mathcal{N}(0,1)$. Toutes ces méthodes reposent sur des simulations de la loi uniforme $\mathcal{U}([0,1])$ obtenues avec la fonction *rand* des ordinateurs.

On note $f_{\mathcal{N}}$ la fonction de densité et $F_{\mathcal{N}}$ la fonction de répartition de la loi $\mathcal{N}(0,1)$. On note également $\mathbf{P}_{\mathcal{N}}$ la loi de probabilité associée.

1. Première méthode

Notons $F_{\mathcal{N}}^{-1}$ la bijection réciproque de $F_{\mathcal{N}}$ définie de $[0,1]$ vers $\mathbf{R}$. On considère une variable X de loi $\mathcal{U}([0,1])$ et on pose $N_1 = F_{\mathcal{N}}^{-1}(X)$.

a. Montrer que N_1 suit la loi normale $\mathcal{N}(0,1)$.

La variable N_1 est à valeurs dans $\mathbf{R}$. Cherchons sa fonction de répartition : soit $t \in \mathbf{R}$.

$$F_{N_1}(t) = \mathbf{P}(N_1 \leqslant t) = \mathbf{P}(F_{\mathcal{N}}^{-1}(X) \leqslant t) = \mathbf{P}(X \leqslant F_{\mathcal{N}}(t)).$$

Comme X suit la loi uniforme $\mathcal{U}([0,1])$ et $F_{\mathcal{N}}(t) \in [0,1]$, cette probabilité est égale à

$$\mathbf{P}_X([0, F_{\mathcal{N}}(t)]) = F_{\mathcal{N}}(t).$$

Ainsi $F_{N_1} = F_{\mathcal{N}}$ et N_1 a la même fonction de répartition que la loi normale. Donc N_1 suit la loi $\mathcal{N}(0,1)$.

b. Pourquoi est-il en pratique difficile de simuler la loi normale de cette façon là ?

Nous ne connaissons pas explicitement la fonction de répartition de la loi normale, c'est-à-dire qu'on ne sait pas l'exprimer à l'aide de fonctions usuelles. Il est donc impossible de déterminer explicitement sa bijection réciproque. On ne peut qu'utiliser des approximations numériques.

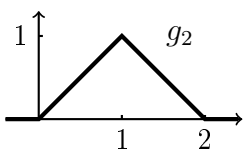
2. Deuxième méthode

Soit $n \geqslant 2$. On note $X_1, \ldots, X_n$ des variables indépendantes de même loi $\mathcal{U}([0,1])$. On pose $S_n = \sum_{i=1}^{n} X_i$ et $N_n = \frac{S_n - nm}{\sigma\sqrt{n}}$ où m et σ désignent l'espérance et l'écart-type des variables X_i.

a. En quoi la variable N_n permet-elle de simuler la loi $\mathcal{N}(0,1)$?

Les variables X_i sont indépendantes et de même loi. Elles ont une espérance $m = \frac{1}{2}$ et une variance $\sigma^2 = \frac{1}{12}$ finies. On peut donc appliquer le théorème central limite : si n est suffisamment grand, la variable N_n suit approximativement la loi $\mathcal{N}(0,1)$. Donc, à partir d'une certaine valeur de n, des simulations de la variable N_n fournissent des simulations presque parfaites de la loi $\mathcal{N}(0,1)$.

On aimerait préciser la valeur de n à partir de laquelle la variable N_n permet de calculer des probabilités pour la loi $\mathcal{N}(0,1)$ avec une précision de 10^{-2}.
La loi de la variable S_n est connue : il s'agit de la loi d'Irwin-Hall. Sa densité g_n (que nous ne donnerons pas ici) permet de déterminer explicitement la densité f_n de N_n.
Le cas $n = 2$ a été étudié en cours et la densité g_2 de S_2 est représentée ci-contre.



b. Calculer à l'aide de cette densité la probabilité $\mathbf{P}(-\sqrt{3/2} \leqslant N_2 \leqslant \sqrt{3/2})$.
La comparer avec la probabilité $\mathbf{P}_{\mathcal{N}}([-\sqrt{3/2}, \sqrt{3/2}])$.

Calculons :

$$\mathbf{P}(-\sqrt{3/2} \leqslant N_2 \leqslant \sqrt{3/2}) = \mathbf{P}(-\sqrt{3/2} \leqslant \frac{S_2 - 2 \cdot \frac{1}{2}}{\sqrt{\frac{2}{12}}} \leqslant \sqrt{3/2}) = \mathbf{P}(\tfrac{1}{2} \leqslant S_2 \leqslant \tfrac{3}{2}) = \int_{\frac{1}{2}}^{\frac{3}{2}} g_2(t)\mathrm{d}t.$$

Cette intégrale se lit facilement graphiquement : elle vaut $\frac{3}{4}$. Donc

$$\mathbf{P}(-\sqrt{3/2} \leqslant N_2 \leqslant \sqrt{3/2}) = \frac{3}{4} = 0,75.$$

Avec la loi normale :

$$\mathbf{P}_{\mathcal{N}}([-\sqrt{3/2}, \sqrt{3/2}]) = 2F_{\mathcal{N}}(\sqrt{3/2}) - 1 \approx 2 \cdot 0,89 - 1 \approx 0,78.$$

L'écart entre les deux résultats est supérieur à $0,01$, la variable N_2 ne permet pas d'obtenir une approximation suffisamment précise de la loi normale.

c. On note $\|f\|_1 = \int_{-\infty}^{+\infty} |f(x)|\mathrm{d}x$. *On reconnaîtra une norme fonctionnelle bien connue.* Justifier que si $\|f_{\mathcal{N}} - f_n\|_1 \leqslant 0,01$, on peut en déduire que pour toute partie A de $\mathbf{R}$,

$$|\mathbf{P}(N_n \in A) - \mathbf{P}_{\mathcal{N}}(A)| \leqslant 0,01.$$

On admet que cette condition est satisfaite à partir de $n = 15$.

On suppose

$$\|f_{\mathcal{N}} - f_n\|_1 = \int_{t \in \mathbf{R}} |f_{\mathcal{N}}(t) - f_n(t)|\mathrm{d}t \leqslant 0,01.$$

Soit A une partie de $\mathbf{R}$. Alors, en notant que $f_{\mathcal{N}}$ et f_n sont positives et en utilisant une inégalité triangulaire,

$$|\mathbf{P}(N_n \in A) - \mathbf{P}_{\mathcal{N}}(A)| = \left|\int_A f_{\mathcal{N}} - \int_A f_n\right| = \left|\int_A f_{\mathcal{N}} - f_n\right| \leqslant \int_A |f_{\mathcal{N}} - f_n| \leqslant \int_{\mathbf{R}} |f_{\mathcal{N}} - f_n|.$$

Donc, d'après l'hypothèse $|\mathbf{P}(N_n \in A) - \mathbf{P}_{\mathcal{N}}(A)| \leqslant 0,01$.

3. Troisième méthode (dite de Box-Muller)

a. Soient X et Y des variables indépendantes de loi normale $\mathcal{N}(0,1)$. Soit B une partie du plan. Exprimer sous forme intégrale la probabilité $\mathbf{P}((X,Y) \in B)$.

Les variables X et Y étant indépendantes, le couple (X,Y) a pour densité $f_X(x)f_Y(y)$. Ces densités f_X et f_Y sont les densités de la loi $\mathcal{N}(0,1)$. Ainsi

$$\mathbf{P}((X,Y) \in B) = \iint_B \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}} \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{y^2}{2}}\mathrm{d}x\mathrm{d}y = \iint_B \frac{1}{2\pi}\mathrm{e}^{-\frac{x^2+y^2}{2}}\mathrm{d}x\mathrm{d}y.$$

b. Effectuer pour cette intégrale un changement de variables en coordonnées polaires :

$$x = r\cos(\theta), \quad y = r\sin(\theta) \quad \text{et} \quad \mathrm{d}x\mathrm{d}y = r\mathrm{d}r\mathrm{d}\theta.$$

Effectuons le changement de variable :

$$\mathbf{P}((X,Y) \in B) = \iint_{\tilde{B}} \frac{1}{2\pi}\mathrm{e}^{-\frac{r^2}{2}} \cdot r\mathrm{d}r\mathrm{d}\theta,$$

où $\tilde{B}$ est l'image de l'ensemble B par le changement de variables.

c. Un exemple : soit B le disque de centre $(0,0)$ et de rayon 1. Que vaut $\mathbf{P}((X,Y)\in B)$?

Ce disque se paramètre en coordonnées polaires par $r\in[0,1]$, $\theta\in[0,2\pi]$. Ainsi

$$\mathbf{P}((X,Y)\in B)=\int_{r=0}^{1}\int_{\theta=0}^{2\pi}\frac{1}{2\pi}re^{-\frac{r^2}{2}}drd\theta=\left[-e^{-\frac{r^2}{2}}\right]_0^1=1-e^{-\frac{1}{2}}\approx 0,39.$$

Les variables r et θ qui apparaissent ci-dessus vont nous permettre de simuler des variables gaussiennes. Considérons maintenant deux variables U_1 et U_2 indépendantes de même loi uniforme $\mathcal{U}([0,1])$.

d. Posons $R=\sqrt{-2\ln(U_1)}$. Déterminer la fonction de répartition de R puis sa densité.

La variable R est à valeurs dans $\mathbf{R}_+^*$. Soit $t>0$. Alors

$$F_R(t)=\mathbf{P}(R\leqslant t)=\mathbf{P}(\sqrt{-2\ln(U_1)}\leqslant t)=\mathbf{P}(U_1\geqslant e^{-\frac{t^2}{2}})=1-e^{-\frac{t^2}{2}}.$$

La densité de R est donnée par

$$f_R(t)=F_R'(t)=te^{-\frac{t^2}{2}}.$$

e. Posons $\Theta=2\pi U_2$. Quelle est la loi de Θ ?

La loi de Θ est simplement la loi uniforme sur $[0,2\pi]$. Sa densité est définie par $f_\Theta(t)=\frac{1}{2\pi}$ pour $t\in[0,2\pi]$.

f. En déduire que $X=R\cos(\Theta)$ et $Y=R\sin(\Theta)$ suivent la loi $\mathcal{N}(0,1)$.

Les variables R et Θ étant indépendantes, la densité du couple (R,Θ) est définie par

$$f_\Theta(\theta)f_R(r)=\frac{1}{2\pi}re^{-\frac{r^2}{2}}.$$

On reconnaît la fonction intégrée dans la question b. après changement de variable. En effectuant le changement de variable dans l'autre sens, on montre bien que la densité du couple $(X,Y)=(R\cos(\Theta),R\sin(\Theta))$ est bien celle d'un couple de variables gaussiennes indépendantes. Ainsi X et Y suivent toutes les deux la loi $\mathcal{N}(0,1)$.

g. En quoi cette méthode est-elle meilleure que la deuxième méthode ?

Avec la seconde méthode, on n'obtient des simulations approximatives de la loi normale. Et pour obtenir des approximations satisfaisantes, il faut commencer par simuler un certain nombre de lois uniformes (au moins 15 pour une précision de l'ordre de 10^{-2}).
Avec la méthode de Box-Muller, on obtient des simulations parfaites de la loi normale et en n'utilisant que deux tirages de la loi uniforme. Si on doit effectuer un grand nombres de simulations, cette technique sera moins coûteuse informatiquement que la seconde.