TD 2 - Autour de la loi normale

Exercice 1. On considère une variable aléatoire X de densité donnée par

$$f(x) = \sqrt{\frac{2}{\pi}} e^{-x^2/2} \mathbb{1}_{]0,\infty[}(x)\,, \quad x \in \mathbb{R}\,.$$

On dit alors que X suit une loi normale tronquée ou loi demi-normale.

1. Proposer une méthode de simulation de X à l'aide de la méthode de rejet avec $Y \sim \mathcal{E}(1)$.

   Notons tout d'abord que f est bien une densité. On note g la densité de la loi exponentielle de paramètre 1 :

   $$g(x) = e^{-x} \mathbb{1}_{[0,\infty[}(x)\,.$$

   On obtient alors

   $$\frac{f(x)}{g(x)} = \sqrt{\frac{2}{\pi}} e^{-\frac{x^2}{2}+x} = \sqrt{\frac{2}{\pi}} e^{-\frac{(x-1)^2-1}{2}} \leqslant \sqrt{\frac{2e}{\pi}}\,.$$

   On pose donc $m = \sqrt{\frac{2e}{\pi}} \approx 1.32$ et $r(x) = e^{-\frac{(x-1)^2}{2}}$. L'algorithme de rejet consiste alors à simuler $Y \sim \mathcal{E}(1)$ et $U \sim \mathcal{U}([0,1])$ et à garder Y si $U \leqslant e^{-\frac{(Y-1)^2}{2}}$.

2. En déduire une méthode de simulation d'une loi normale.

   La question précédente permet de simuler une loi normale sur l'axe des réels positifs. Il reste à attribuer un signe à une telle réalisation. L'idée est de choisir aléatoirement et indépendamment le signe.
   On considère une variable aléatoire B de loi de Bernoulli de paramètre $1/2$ indépendante de X simulée selon f avec la question précédente. On pose alors $N = BX - (1-B)X$ qui vaut donc X si B vaut 1 et $-X$ si B vaut 0. Montrons que N suit une loi normale.
   Pour $x \in \mathbb{R}$, la formule des probabilités totales donne

   $$\begin{aligned}\mathbb{P}(N \leqslant x) &= \mathbb{P}(N \leqslant x \mid B=1)\mathbb{P}(B=1) + \mathbb{P}(N \leqslant x \mid B=0)\mathbb{P}(B=0)\\ &= \mathbb{P}(X \leqslant x \mid B=1)\frac{1}{2} + \mathbb{P}(-X \leqslant x \mid B=0)\frac{1}{2}\\ &= \frac{\mathbb{P}(X \leqslant x) + \mathbb{P}(-X \leqslant x)}{2}\,,\end{aligned}$$

   où on a utilisé l'indépendance de X et B dans la deuxième égalité. Par suite, si $x \leqslant 0$, alors la probabilité précédente devient

   $$\frac{\mathbb{P}(-X \leqslant x)}{2} = \frac{\mathbb{P}(X \geqslant -x)}{2} = \int_{-x}^{\infty} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} \mathrm{d}t = \int_{-\infty}^{x} \frac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}} \mathrm{d}u\,,$$

   où la dernière égalité résulte du changement de variable $u = -t$. De même, pour $x > 0$, on obtient

   $$\mathbb{P}(N \leqslant x) = \frac{\mathbb{P}(X \leqslant x) + 1}{2} = \int_0^x \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} \mathrm{d}t + \frac{1}{2} = \int_{-\infty}^{x} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} \mathrm{d}t\,,$$

   où on a utilisé ici l'égalité $\int_{-\infty}^{0} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} = \frac{1}{2}$. Ainsi, pour tout $x \in \mathbb{R}$, on a

   $$\mathbb{P}(N \leqslant x) = \int_{-\infty}^{x} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} \mathrm{d}t\,,$$

   ce qui prouve que N suit une loi normale centrée réduite.

Exercice 2. Une variable aléatoire suit une loi de Laplace de paramètre $\lambda > 0$ si sa densité est donnée par

$$g(x) = \frac{\lambda}{2} e^{-\lambda|x|}, \quad x \in \mathbb{R}.$$

1. Vérifier que cela définit bien une loi de probabilité.

   La fonction g est mesurable, positive et vérifie

   $$\int_{\mathbb{R}} g(x)\,\mathrm{d}x = \frac{\lambda}{2}\left(\int_{-\infty}^{0} e^{\lambda x}\,\mathrm{d}x + \int_{0}^{+\infty} e^{-\lambda x}\,\mathrm{d}x\right) = \frac{1}{2}\left(\left[e^{\lambda x}\right]_{-\infty}^{0} - \left[e^{\lambda x}\right]_{0}^{+\infty}\right) = 1.$$

   Ainsi, g est bien une densité.

2. À l'aide de la méthode de rejet, proposer une méthode pour simuler une loi normale centrée réduite à partir de g. Pour quelle valeur de λ la probabilité de rejet est-elle minimale ?

   On procède comme dans l'exercice précédent, avec cette fois-ci

   $$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}.$$

   On obtient

   $$\frac{f(x)}{g(x)} = \frac{2}{\lambda\sqrt{2\pi}} e^{-\frac{x^2}{2}+\lambda|x|} = \frac{1}{\lambda}\sqrt{\frac{2}{\pi}} e^{-\frac{(|x|-\lambda)^2+\lambda}{2}} \leqslant \frac{1}{\lambda}\sqrt{\frac{2}{\pi}} e^{\frac{\lambda^2}{2}} =: m(\lambda).$$

   Il reste à minimiser $m(\lambda)$ ce qui revient à minimiser son logarithme. La dérivée de $\lambda \mapsto \ln(m(\lambda))$ vaut $-\frac{1}{\lambda} + \lambda$ donc s'annule en $\lambda = 1$. Le minimum vaut alors $m(1) = \sqrt{\frac{2e}{\pi}}$. On retrouve la borne de l'exercice précédent.
   L'algorithme de rejet consiste alors à simuler une variable aléatoire $U \sim \mathcal{U}([0,1])$ et Y selon g, puis à conserver Y si $U \leqslant e^{-\frac{(|Y|-1)^2}{2}}$. Remarquons que Y se simule sans problème avec la méthode d'inversion (sa fonction de répartition est bijective).
   D'après le cours, la probabilité de rejet est donnée par

   $$1 - \frac{1}{m} = 1 - \sqrt{\frac{\pi}{2e}} \approx 0.24.$$

Exercice 3. On considère deux variables aléatoires U et V indépendantes de loi uniforme sur $[0,1]$.

1. Rappeler le principe de la méthode de Box-Müller.

   La méthode de Box-Müller consiste à poser

   $$N_1 = \sqrt{-2\ln(U)}\cos(2\pi V) \quad \text{et} \quad N_2 = \sqrt{-2\ln(U)}\sin(2\pi V).$$

   Le théorème de changement de variables en coordonnées polaires assure alors que N_1 et N_2 sont deux variables aléatoires indépendantes de loi normale centrée réduite.

2. Rappeler comment simuler un couple (X, Y) de loi uniforme sur le disque unité.

   Il suffit de considérer le couple $(2U-1, 2V-1)$ qui suit une loi uniforme sur le pavé $[-1,1]^2$. Ensuite, on pose $(X,Y) = (2U-1, 2V-1)$ si le vecteur appartient au disque, sinon on répète l'opération. Voir l'exercice 8 du TD 1 pour davantage de détails.

3. On pose $T = X^2+Y^2$. Montrer que le couple $\left(X\sqrt{-2\ln(T)/T}, Y\sqrt{-2\ln(T)/T}\right)$ est formée de deux variables aléatoires indépendantes de loi normales centrées réduites. On pourra étudier la loi des composantes radiale et angulaire de (X,Y).
   Remarque : cette méthode de simulation est due à Marsaglia.

   La composante radiale de (X,Y) est la variable aléatoire $T = X^2+Y^2$ qui vérifie
   $$\mathbb{P}(T \leqslant t) = \mathbb{P}(\sqrt{X^2+Y^2} \leqslant \sqrt{t}) = \mathbb{P}((X,Y) \in \mathcal{D}_{\sqrt{t}}), \quad t \geqslant 0,$$
   où $\mathcal{D}_{\sqrt{t}}$ désigne le disque de rayon $\sqrt{t}$. Comme le vecteur (X,Y) suit une loi uniforme sur le disque $\mathcal{D}$ on en déduit que
   $$\mathbb{P}(T \leqslant t) = \frac{\text{Aire}(\mathcal{D}_{\sqrt{t}})}{\text{Aire}(\mathcal{D})} = \frac{\sqrt{t}^2\pi}{1^2\pi} = t.$$
   Ainsi, T suit une loi uniforme sur $[0,1]$.
   D'autre part, notons Θ la variable aléatoire aléatoire donnant l'angle du vecteur (X,Y), c'est-à-dire
   $$\cos(\Theta) = \frac{X}{X^2+Y^2} \quad \text{et} \quad \sin(\Theta) = \frac{Y}{X^2+Y^2}.$$
   Cette quantité vérifie la relation
   $$\mathbb{P}(\Theta \leqslant \alpha) = \mathbb{P}((X,Y) \in \mathcal{A}_\alpha),$$
   où $\mathcal{A}_\alpha$ désigne l'ensemble des points du disque unité formant un angle inférieur à α avec l'axe des abscisses. Ce secteur angulaire a pour superficie $\frac{\alpha}{2}$, ce qui donne
   $$\mathbb{P}(\Theta \leqslant \alpha) = \mathbb{P}((X,Y) \in \mathcal{A}_\alpha) = \frac{\frac{\alpha}{2}}{\pi} = \frac{\alpha}{2\pi}.$$
   On a utilisé la formule $\text{Aire}(\mathcal{A}_{\alpha,R}) = \frac{\alpha}{2}R^2$, où $\mathcal{A}_{\alpha,R}$ désigne l'ensemble des points du disque de rayon R d'angle inférieur à α.
   Bref, le couple proposé s'écrit
   $$\left(X\sqrt{\frac{-2\ln(T)}{T}}, Y\sqrt{\frac{-2\ln(T)}{T}}\right) = \left(\sqrt{-2\ln(T)}\cos(\Theta), \sqrt{-2\ln(T)}\sin(\Theta)\right)$$
   et a donc la même loi que celui proposé par l'algorithme de Box-Müller. Les deux marginales suivent ainsi une loi normale centrée réduite.

4. Comparer cette approche avec la méthode de Box-Müller.

   Cette approche a l'avantage de ne pas faire intervenir de fonctions trigonométriques (coûteuses en temps de calcul), mais s'appuie sur une procédure de rejet. Rappelons que la probabilité de rejet pour simuler des variables aléatoires sur le disque est de 21% (voir l'exercice 8 du TD 1), ce qui est assez faible.

Exercice 4. Soit $X_1, \ldots, X_n$ des variables aléatoires iid de loi $\mathcal{N}(\mu, \sigma^2)$. Un estimateur de μ (resp. σ^2) est une quantité aléatoire basée sur l'échantillon $X_1, \ldots, X_n$ qui permet d'approcher μ (resp. σ^2). Le but de l'exercice est de proposer des estimateurs naturels de μ et σ^2 et d'étudier leurs propriétés.

1. Cas 1 : μ inconnue et σ connu
   On suppose que la valeur de μ est inconnue mais que celle de σ est connue.

(a) Proposer un estimateur $\bar{X}_n$, basé sur l'échantillon $X_1, \dots, X_n$, permettant d'approcher μ. Calculer l'espérance de cet estimateur et donner sa limite (en un sens à préciser) quand $n \to \infty$.

Un estimateur "naturel" de l'espérance μ est la moyenne empirique

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}\,.$$

L'espérance de cet estimateur vaut

$$\mathrm{E}[\bar{X}_n] = \frac{\mathrm{E}[X_1 + \cdots + X_n]}{n} = \frac{\mathrm{E}[X_1] + \cdots + \mathrm{E}[X_n]}{n} = \mu\,.$$

Par ailleurs, la loi des grands nombres assure la convergence presque sûre de $\bar{X}_n$ vers μ. Notons qu'ici on n'a pas utilisé le caractère gaussien des X_i.

(b) Donner la loi de $\bar{X}_n$.

Les variables aléatoires $X_1, \dots, X_n$ sont indépendantes et de loi normale donc leur somme suit encore une loi normale. De plus, la variance de $\bar{X}_n$ est donnée par

$$\mathrm{Var}(\bar{X}_n) = \mathrm{Var}\Big(\frac{X_1 + \cdots + X_n}{n}\Big) = \frac{1}{n}\Big(\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)\Big) = \frac{\sigma^2}{n}\,.$$

Ainsi,

$$\bar{X}_n \sim \mathcal{N}\Big(\mu, \frac{\sigma^2}{n}\Big)\,.$$

(c) Déterminer un intervalle $I_1(X_1, \dots, X_n)$ qui contient μ avec probabilité $1 - \alpha$, pour $\alpha > 0$.

À partir de la question précédente, on en déduit que $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ suit une loi normale centrée réduite. Notons q_z le quantile d'ordre z de la loi normale, c'est-à-dire vérifiant $\mathbb{P}(N \leqslant q_z) = z$, avec $N \sim \mathcal{N}(0,1)$. On obtient alors par symétrie de la loi normale que

$$\mathbb{P}(-q_{1-\frac{\alpha}{2}} \leqslant \sqrt{n}(\bar{X}_n - \mu)/\sigma \leqslant q_{1-\frac{\alpha}{2}}) = 1 - \alpha\,.$$

ce qui se réécrit

$$\mathbb{P}\Big(\bar{X}_n - \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}} \leqslant \mu \leqslant \bar{X}_n + \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}}\Big) = 1 - \alpha\,.$$

Ainsi, l'intervalle $I_1(X_1, \dots, X_n)$, donné par

$$I_1(X_1, \dots, X_n) = \Big[\bar{X}_n - \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}}\Big]\,,$$

contient μ avec une probabilité de $1 - \alpha$.

2. Cas 2 : σ inconnu et μ connu

On suppose que la valeur de σ est inconnue mais que celle de μ est connue. On rappelle que pour une variable aléatoire de loi normale centrée réduite X, les moments sont donnés par

$$\mathrm{E}[X^{2p+1}] = 0 \quad \text{et} \quad \mathrm{E}[X^{2p}] = \frac{(2p)!}{2^p p!}\,, \quad p \in \mathbb{N}\,.$$

(a) Proposer un estimateur $\hat{\sigma}_n^2$, basé sur l'échantillon $X_1, \dots, X_n$, permettant d'approcher σ^2. Calculer l'espérance et la variance de cet estimateur.

Rappelons que $\sigma^2 = \mathrm{Var}(X_1) = \mathrm{E}[(X_1 - \mu)^2]$. Un estimateur "naturel" de σ^2 est alors

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{k=1}^{n} (X_k - \mu)^2 \,.$$

L'espérance de $\hat{\sigma}_n$ vaut alors

$$\mathrm{E}[\hat{\sigma}_n^2] = \frac{1}{n} \sum_{k=1}^{n} \mathrm{E}[(X_k - \mu)^2] = \mathrm{Var}(X_1) = \sigma^2 \,.$$

Pour la variance on utilise la formule de Koenig-Huygens :

$$\mathrm{Var}(\hat{\sigma}_n^2) = \frac{1}{n} \mathrm{Var}((X_1 - \mu)^2) = \frac{1}{n} \Big(\mathrm{E}[(X_1 - \mu)^4] - \mathrm{E}[(X_1 - \mu)^2]^2 \Big) \,.$$

à partir de la la formule donnée dans l'énoncé on obtient

$$\mathrm{E}[(X_1 - \mu)^4] = \sigma^4 \mathrm{E}\Big[\Big(\frac{X_1 - \mu}{\sigma} \Big)^4 \Big] = \sigma^4 \mathrm{E}[N^4] = \sigma^4 \frac{4!}{2^2\, 2!} = 3\sigma^4 \,,$$

où N désigne une variable aléatoire de loi normale centrée réduite. Par ailleurs, le terme $\mathrm{E}[(X_1 - \mu)^2]^2$ correspond au carré de la variance de X_1, soit σ^4. On en déduit alors l'expression de la variance de $\hat{\sigma}_n^2$:

$$\mathrm{Var}(\hat{\sigma}_n^2) = \frac{2\sigma^4}{n} \,.$$

(b) Quelle est la loi de $\frac{n\hat{\sigma}_n^2}{\sigma^2}$? En déduire un intervalle $I_2(X_1, \dots, X_n)$ qui contient σ avec probabilité $1 - \alpha$, pour $\alpha > 0$.

On remarque que la variable aléatoire

$$\frac{n\hat{\sigma}_n^2}{\sigma^2} = \sum_{j=1}^{n} \Big(\frac{X_j - \mu}{\sigma} \Big)^2$$

est une somme de n variables aléatoires indépendantes de loi normale centrée réduite. Ainsi, $\frac{n\hat{\sigma}_n^2}{\sigma^2}$ suit une loi du chi-deux à n degrés de liberté.
Notons que la loi du chi-deux n'est pas symétrique : elle prend ses valeurs dans $[0, \infty[$ (ce qui est rassurant car $\hat{\sigma}_n^2$ est positif). Notons $q'_{\frac{\alpha}{2}}$ (resp. $q'_{1-\frac{\alpha}{2}}$) le quantile d'ordre $\frac{\alpha}{2}$ (resp. $1 - \frac{\alpha}{2}$) de la loi du chi-deux à n degrés de liberté. Alors,

$$\mathbb{P}\Big(q'_{\frac{\alpha}{2}} \leqslant \frac{n\hat{\sigma}_n^2}{\sigma^2} \leqslant q'_{1-\frac{\alpha}{2}} \Big) = 1 - \alpha \,.$$

On prendra garde au fait que $q'_{\frac{\alpha}{2}}$ et $q'_{1-\frac{\alpha}{2}}$ ne sont pas opposés l'un de l'autre car la loi du chi-deux n'est pas symétrique. On obtient alors l'intervalle de confiance (asymétrique)

$$I_2(X_1, \dots, X_n) = \Big[\sqrt{\frac{n}{q'_{1-\frac{\alpha}{2}}}} \hat{\sigma}_n,\ \sqrt{\frac{n}{q'_{\frac{\alpha}{2}}}} \hat{\sigma}_n, \Big] \,.$$

3. Cas 3 : μ et σ inconnus

   On suppose que l'espérance μ et l'écart-type σ sont tous les deux inconnus.

   (a) Peut-on toujours utiliser $\bar{X}_n$ comme estimateur de μ ? Et $\hat{\sigma}_n^2$ comme estimateur de σ^2 ?

   La quantité $\bar{X}_n$ ne dépend que des X_i donc peut toujours être utilisée comme estimateur de μ. Par contre, σ_n^2 dépend de μ qui est inconnu, donc on ne peut plus utiliser cette quantité pour estimer σ^2.

   (b) On choisit comme estimateur de σ^2 la quantité

$$S_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2 .$$

   On peut montrer que la variable aléatoire $\frac{nS_n^2}{\sigma^2}$ suit une loi du chi-deux à $n-1$ degrés de liberté (on notera le perte d'un degré de liberté par rapport au cas précédent) et que cette variable aléatoire est indépendante de $\bar{X}_n$ (cela semble étonnant mais ce résultat peut se montrer grâce au théorème de Cochran vu en M1).
   Donner la loi de

$$\sqrt{n-1}\frac{\bar{X}_n - \mu}{S_n} .$$

   En déduire un intervalle $I_3(X_1, \ldots, X_n)$ qui contient μ avec probabilité $1-\alpha$, pour $\alpha > 0$.

   On part de l'égalité

$$\sqrt{n-1}\frac{\bar{X}_n - \mu}{S_n} = \frac{\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}}{\sqrt{\frac{nS_n^2}{\sigma^2}} \times \frac{1}{\sqrt{n-1}}} .$$

   Cette quantité correspond au ratio d'une loi normale par une loi du chi-deux à $n-1$ degrés de liberté, les deux variables aléatoires étant indépendantes d'après l'énoncé. Le cours assure alors que $\sqrt{n-1}\frac{\bar{X}_n - \mu}{S_n}$ suit une loi de Student à $n-1$ degrés de liberté.
   En notant $\tilde{q}_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\frac{\alpha}{2}$ d'une loi de Student (qui est symétrique), on en déduit l'intervalle de confiance demandé :

$$I_3(X_1, \ldots, X_n) = \left[\bar{X}_n - \frac{S_n}{\sqrt{n-1}}\tilde{q}_{1-\frac{\alpha}{2}}, \bar{X}_n + \frac{S_n}{\sqrt{n-1}}\tilde{q}_{1-\frac{\alpha}{2}}\right] .$$