

## PC 2 (Modèle statistique)

---

### 1 Modèle exponentiel

Une grande partie des modèles utilisés dans les exemples élémentaires sont des modèles exponentiels (modèle gaussien, log-normal, exponentiel, gamma, Bernouilli, Poisson, etc). Nous allons étudier quelques propriétés de ces modèles. On appelle modèle exponentiel une famille de lois  $\{\mathbb{P}_\theta, \theta \in \Theta\}$  ayant une densité par rapport à une mesure  $\mu$   $\sigma$ -finie sur  $\mathbb{R}$  ou  $\mathbb{N}$  de la forme

$$p_\theta(x) = c(\theta) \exp(m(\theta)f(x) + h(x)).$$

On supposera que  $\Theta$  est un intervalle ouvert de  $\mathbb{R}$ ,  $m(\theta) = \theta$ ,  $c$  de classe  $C^2$ ,  $c(\theta) > 0$  pour tout  $\theta \in \Theta$ . On notera  $X$  une variable aléatoire de loi  $\mathbb{P}_\theta$  et on admettra que

$$\frac{\partial^i}{\partial \theta^i} \int \exp(\theta f(x) + h(x)) \mu(dx) = \int f(x)^i \exp(\theta f(x) + h(x)) \mu(dx) < +\infty, \quad \text{pour } i = 1, 2.$$

1. Montrez que  $\varphi(\theta) := \mathbb{E}_\theta(f(X)) = -\frac{d}{d\theta} \log(c(\theta))$ .
2. Montrez que  $\text{Var}_\theta(f(X)) = \varphi'(\theta) = -\frac{d^2}{d\theta^2} \log(c(\theta))$ .
3. On dispose d'un  $n$ -échantillon  $X_1, \dots, X_n$  de loi  $\mathbb{P}_\theta$ . On note  $\hat{\theta}_n$  l'estimateur obtenu en résolvant  $\varphi(\hat{\theta}_n) = \frac{1}{n} \sum_{i=1}^n f(X_i)$ . En supposant  $\text{Var}_\theta(f(X)) > 0$ , montrez que

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\text{loi}} \mathcal{N}\left(0, \frac{1}{\text{Var}_\theta(f(X))}\right).$$

#### Corrigé :

Observer que puisque  $p_\theta$  est une densité, on a

$$\frac{1}{c(\theta)} = \int \exp(\theta f(x) + h(x)) \mu(dx),$$

et que le résultat admis dit que  $\theta \mapsto 1/c(\theta)$  est de classe  $C^2$ .

1. D'après le résultat admis,

$$\int f(x) p_\theta(x) \mu(dx) = c(\theta) \frac{d}{d\theta} \left( \frac{1}{c(\theta)} \right)$$

ce qui donne le résultat demandé.

2. De même, on a

$$\int f^2(x) p_\theta(x) \mu(dx) = c(\theta) \frac{d^2}{d\theta^2} \frac{1}{c(\theta)}$$

ce qui donne

$$\int f^2(x) p_\theta(x) \mu(dx) = -\frac{c''(\theta)}{c(\theta)} + 2 \left( \frac{c'(\theta)}{c(\theta)} \right)^2$$

En combinant ce calcul avec la question précédente, on obtient

$$\text{Var}_\theta(f(X)) = \left( \frac{c'(\theta)}{c(\theta)} \right)^2 - \frac{c''(\theta)}{c(\theta)}$$

ce qui donne le résultat demandé.

3. Par le TCL pour des v.a. i.i.d.

$$\sqrt{n} \left( \phi(\hat{\theta}_n) - \mathbb{E}_\theta[f(X)] \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \text{Var}_\theta(f(X)))$$

puis on conclut par la méthode delta appliquée avec la fonction  $\phi^{-1}$ .

## 2 Estimation par la méthode plug-in

Soit  $X_1, \dots, X_n$  des variables aléatoires réelles i.i.d. de fonction de répartition  $F$ , soit  $a < b$  deux réels et soit  $\theta = F(b) - F(a)$ .

1. Déterminer l'estimateur plug-in  $\hat{\theta}$  de  $\theta$ .
2. Déterminer l'estimateur plug-in de la variance de  $\hat{\theta}$  et en déduire un intervalle de confiance asymptotique pour  $\theta$  de niveau  $1 - \alpha$ .

**Corrigé :**

1. L'estimateur de substitution est donné par

$$\hat{\theta}_n = \hat{F}_n(b) - \hat{F}_n(a) = \frac{1}{n} \sum_{k=1}^n \mathbf{1}_{a < X_k \leq b}.$$

2. Notons  $\hat{\sigma}_n^2$  l'estimateur plug-in de la variance de  $\hat{\theta}_n$ . On a

$$\hat{\sigma}_n^2 = \frac{\hat{\theta}_n (1 - \hat{\theta}_n)}{n}.$$

Le TCL pour des v.a. i.i.d. et le lemme de Slutsky (noter que  $n\hat{\sigma}_n^2 \xrightarrow{\text{p.s.}} \theta(1-\theta)$ ) donnent

$$\frac{1}{\hat{\sigma}_n} (\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Notons  $z_{1-\alpha/2}$  le quantile d'ordre  $1 - \alpha/2$  d'une loi  $\mathcal{N}(0, 1)$ . On déduit de cette convergence en loi un intervalle de confiance asymptotique pour  $\theta$  au niveau  $(1 - \alpha)$

$$\left[ \hat{\theta}_n - z_{1-\alpha/2} \hat{\sigma}_n; \hat{\theta}_n + z_{1-\alpha/2} \hat{\sigma}_n \right]$$

## 3 Stabilisation de la variance

On dispose d'un échantillon  $X_1, \dots, X_n$  i.i.d. de loi de Bernoulli de paramètre  $0 < \theta < 1$ .

1. On note  $\bar{X}_n$  la moyenne empirique des  $X_i$ . Que disent la loi des grands nombres et le TCL ?
2. Cherchez une fonction  $g$  telle que  $\sqrt{n}(g(\bar{X}_n) - g(\theta))$  converge en loi vers  $Z$  de loi  $\mathcal{N}(0, 1)$ .
3. On note  $z_\alpha$  le quantile d'ordre  $1 - \alpha/2$  de la loi normale standard. En déduire un intervalle  $\hat{I}_{n,\alpha}$  fonction de  $z_\alpha, n, \bar{X}_n$  tel que  $\lim_{n \rightarrow \infty} \mathbb{P}(\theta \in \hat{I}_{n,\alpha}) = 1 - \alpha$ .

**Corrigé :**

1. On a

$$\bar{X}_n \xrightarrow{\text{p.s.}} \mathbb{E}_\theta[X] = \theta \quad \sqrt{n} (\bar{X}_n - \mathbb{E}_\theta[X]) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \theta(1 - \theta)).$$

2. Par la méthode delta, on cherche une fonction  $g$  dérivable en  $\theta$ , pour tout  $\theta \in ]0, 1[$ , telle que  $(g'(\theta))^2 \theta(1 - \theta) = 1$ . Plusieurs solutions possibles

$$\theta \mapsto \arccos(2\theta - 1) \qquad \theta \mapsto 2 \arccos \sqrt{\theta}$$

et puis les analogues avec arcsin puisque  $\arcsin(u) + \arccos(u) = \pi/2$  pour tout  $u \in [-1, 1]$ .

3. On en déduit l'intervalle de confiance asymptotique de niveau  $1 - \alpha$

$$I_{n,\alpha} = \left[ \frac{1}{2} \left( 1 + \cos \left( g(\overline{X}_n) + \frac{z_\alpha}{\sqrt{n}} \right) \right); \frac{1}{2} \left( 1 + \cos \left( g(\overline{X}_n) - \frac{z_\alpha}{\sqrt{n}} \right) \right) \right]$$

avec  $g(x) = \arccos(2x - 1)$ .

## 4 Modèle d'autorégression

On considère l'observation  $Z = (X_1, \dots, X_n)$ , où les  $X_i$  sont issus du processus d'autorégression :

$$X_i = \theta X_{i-1} + \xi_i, \quad i = 1, \dots, n, \quad X_0 = 0,$$

avec les  $\xi_i$  i.i.d. de loi normale  $\mathcal{N}(0, \sigma^2)$  et  $\theta \in \mathbb{R}$ . Écrire le modèle statistique engendré par l'observation  $Z$ .

**Corrigé :**

- espace probabilisable :  $\mathbb{R}^n$  muni de la tribu borélienne.
- une famille de lois  $\mathbb{P}_\psi$  sur  $\mathbb{R}^n$  données par

$$\mathbb{P}_\psi(A) = \int_A \frac{1}{(\sqrt{2\pi}\sigma)^n} \frac{1}{\sqrt{\det(\Gamma)}} \exp \left( -\frac{1}{2\sigma^2} \underline{x}^T \Gamma^{-1} \underline{x} \right) d\underline{x}$$

où  $\psi = (\theta, \sigma) \in \mathbb{R} \times ]0, \infty[$  et

$$\Gamma^{-1} = \begin{bmatrix} 1 + \theta^2 & -\theta & 0 & \dots & \dots \\ -\theta & 1 + \theta^2 & -\theta & 0 & \dots \\ 0 & -\theta & 1 + \theta^2 & -\theta & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & -\theta & 1 + \theta^2 & -\theta \\ \dots & \dots & 0 & -\theta & 1 \end{bmatrix}$$

Noter que  $\Gamma = A A^T$  avec

$$A = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ \theta & 1 & \dots & 0 & 0 \\ \theta^2 & \theta & 1 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \theta^{n-1} & \theta^{n-2} & \dots & \theta & 1 \end{bmatrix}$$

Pour obtenir l'expression de la loi  $\mathbb{P}_\psi$ , on peut remarquer que  $X_k$  est une combinaison linéaire des v.a.  $\xi_1, \dots, \xi_k$ ; et écrire le vecteur  $Z$  comme une transformation affine du vecteur gaussien  $(\xi_1, \dots, \xi_n)$ .

**Autre expression** L'expression de  $\mathbb{P}_\psi(A)$  est équivalente à

$$\mathbb{P}_\psi(A) = \int_A \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta x_{i-1})^2 \right) d\underline{x}$$

que l'on obtient par exemple de la façon suivante. Soit  $f$  la densité de probabilité du vecteur  $(\xi_1, \xi_1, \dots, \xi_n)$  :

$$f(z_1, z_1, \dots, z_n; \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n z_i^2\right)$$

Soit  $\phi$  l'application de  $\mathbb{R}^n$  dans  $\mathbb{R}^n$  définie par

$$\phi(x_1, x_2, \dots, x_n) = (z_1, z_1, \dots, z_n)$$

Le Jacobien de  $\phi$  est la matrice

$$J_\phi = \begin{pmatrix} 1 & -\theta & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & -\theta & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & -\theta & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & 1 & -\theta \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

On vérifie facilement que son déterminant vaut 1. Le déterminant du Jacobien de  $\phi^{-1}$  vaut donc également 1. On en déduit l'expression de la densité  $g$  de  $(X_1, \dots, X_n)$  :

$$\begin{aligned} g(x_1, x_2, \dots, x_n; \theta, \sigma^2) &= f(\phi(x_1, x_2, \dots, x_n; \sigma^2)) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta x_{i-1})^2\right) \end{aligned}$$

## 5 Survie

On étudie un système qui fonctionne si deux machines de types différents fonctionnent. Les durées de vie  $X_1$  et  $X_2$  des deux machines suivent des lois exponentielles de paramètres  $\lambda_1$  et  $\lambda_2$  :  $\mathbb{P}(X_i > x) = e^{-\lambda_i x}$ . Les variables aléatoires  $X_1$  et  $X_2$  sont supposées indépendantes.

1. Calculer la probabilité pour que le système ne tombe pas en panne avant la date  $t$ . En déduire la loi de la durée de vie  $Z$  du système. Calculer la probabilité pour que la panne du système soit due à une défaillance de la machine 1.
2. Soit  $I = 1$  si la panne du système est due à une défaillance de la machine 1,  $I = 0$  sinon. Calculer  $\mathbb{P}(Z > t; I = \delta)$ , pour tout  $t \geq 0$  et  $\delta \in \{0, 1\}$ . En déduire que  $Z$  et  $I$  sont indépendantes.
3. On dispose de  $n$  systèmes identiques et fonctionnant indépendamment les uns des autres dont on observe les durées de vie  $Z_1, \dots, Z_n$ .
  - (a) Écrire le modèle statistique correspondant. Les paramètres  $\lambda_1$  et  $\lambda_2$  sont-ils identifiables ?
  - (b) Supposons maintenant que l'on observe à la fois les durées de vie des systèmes  $Z_1, \dots, Z_n$  et les causes de la défaillance correspondantes  $I_1, \dots, I_n$ ,  $I_i \in \{0, 1\}$ . Écrire le modèle statistique dans ce cas. Les paramètres  $\lambda_1$  et  $\lambda_2$  sont-ils identifiables ?

**Corrigé :**

1. La probabilité que le système ne tombe pas en panne avant la date  $t$  est

$$\mathbb{P}(Z > t) = \exp(-(\lambda_1 + \lambda_2)t).$$

La probabilité que la panne soit due à la défaillance de la machine 1 est

$$\mathbb{P}(Z = X_1) = \mathbb{P}(X_2 \geq X_1) = \frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

2. Pour tout  $t \geq 0$ ,

$$\mathbb{P}(Z > t, I = 1) = \mathbb{P}(X_2 \geq X_1, X_1 > t) = \frac{\lambda_1}{\lambda_1 + \lambda_2} \exp(-(\lambda_1 + \lambda_2)t) = \mathbb{P}(Z = X_1) \mathbb{P}(Z > t).$$

Résultat analogue pour  $\mathbb{P}(Z > t, I = 0)$ .

3. (a) Modèle statistique

- Espace probabilisable :  $]0, +\infty[^n$  muni de la tribu borélienne.
- Famille de fonctions de répartition  $F_\theta$  sur  $]0, \infty[$  indexées par  $\theta = (\lambda_1, \lambda_2) \in \Theta = ]0, +\infty[^2$  et définies par

$$F_\theta(x) = (1 - \exp(-(\lambda_1 + \lambda_2)x)) \mathbf{1}_{\mathbb{R}^+}(x)$$

Modèle non identifiable.

- (b) Modèle statistique

- Espace probabilisable :  $(]0, +\infty[ \times \{0, 1\})^n$  muni de la tribu engendrée par les ensembles  $A \times \{0\}, A \times \{1\}$  où  $A$  est un borélien.
- Famille de fonctions de répartition  $F_\theta$  sur  $]0, \infty[ \times \{0, 1\}$  indexées par  $\theta = (\lambda_1, \lambda_2) \in \Theta = ]0, +\infty[^2$  et définies par

$$F_\theta(x, 1) = (1 - \exp(-(\lambda_1 + \lambda_2)x)) \mathbf{1}_{\mathbb{R}^+}(x) \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

$$F_\theta(x, 0) = (1 - \exp(-(\lambda_1 + \lambda_2)x)) \mathbf{1}_{\mathbb{R}^+}(x) \frac{\lambda_2}{\lambda_1 + \lambda_2}$$

Modèle identifiable.

---

## Exercices bonus

**Exercice 1.** Pour tout  $\alpha > 0$ , on appelle loi Gamma( $\alpha$ ) la loi sur  $\mathbb{R}^+$  de densité

$$g_\alpha(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}, \quad \text{où } \Gamma(\alpha) \triangleq \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

Pour  $a, b > 0$ , on appelle loi Beta( $a, b$ ) la loi sur  $[0, 1]$  de densité

$$h_{a,b}(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1}.$$

1. Soit  $s$  et  $t > 0$  et soit  $X$  et  $Y$  deux variables indépendantes de loi Gamma( $s$ ) et Gamma( $t$ ), respectivement. On pose

$$\begin{aligned} U &= X + Y \\ V &= X/(X + Y) \end{aligned}$$

Montrer que  $U$  et  $V$  sont indépendantes et que  $U$  est distribuée suivant une loi Gamma( $s+t$ ) et  $V$  suivant une loi Beta( $s, t$ ). [Indication : on pourra considérer la densité jointe de  $(U, V)$  sans se préoccuper des constantes de normalisation.]

2. Soit  $\{Z_n\}_{n \geq 0}$  une suite de variables aléatoires telles que, pour tout  $n \geq 0$ ,  $Z_n$  est de loi Gamma( $n$ ). Montrer que

$$\sqrt{n} \left( \frac{Z_n}{n} - 1 \right) \xrightarrow{(\text{loi})} \mathcal{N}(0, 1) .$$

3. Soient  $p \in (0, 1)$  et  $\{k_n\}$  une suite monotone croissante d'entiers vérifiant

$$\sqrt{n} \left( \frac{k_n}{n} - p \right) \rightarrow 0 . \quad (1)$$

Soient  $\{X_n\}_{n \geq 0}$  et  $\{Y_n\}_{n \geq 0}$  deux suites indépendantes telles que  $X_n \sim \text{Gamma}(k_n)$  et  $Y_n \sim \text{Gamma}(n - k_n)$ . On pose

$$V_n = \frac{X_n}{X_n + Y_n} .$$

Montrer que

$$\sqrt{n} (V_n - p) \xrightarrow{(\text{loi})} \mathcal{N}(0, p(1 - p)) .$$

[Indication : on pourra, dans un premier temps, considérer le comportement asymptotique du couple  $\frac{1}{n}(X_n, Y_n) - (p, 1 - p)$ .]

4. Conclure.

### Corrigé :

1. Les v.a.  $X, Y$  étant indépendantes, la loi jointe est donnée par

$$f_{(X,Y)}(x, y) = f_X(x) f_Y(y) \propto x^{s-1} \exp(-x) y^{t-1} \exp(-y) \mathbf{1}_{\mathbb{R}^+}(x) \mathbf{1}_{\mathbb{R}^+}(y) .$$

Par le changement de variable

$$\phi : \begin{bmatrix} x \\ y \end{bmatrix} \rightarrow \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} x + y \\ \frac{x}{x+y} \end{bmatrix}$$

on obtient pour toute fonction  $h$  mesurable positive

$$\mathbb{E}[h(U, V)] \propto \int_{\mathbb{R}_+^+ \times ]0,1[} h(u, v) u^{s+t-1} \exp(-u) v^{s-1} (1-v)^{t-1} du dv$$

dont on déduit que  $(U, V)$  sont deux variables indépendantes ;  $U$  est une loi Gamma de paramètre  $s + t$  et  $V$  est une loi Beta de paramètres  $(s, t)$ .

2. Par la question précédente,  $Z_n$  a même loi que  $\sum_{k=1}^n W_k$  où  $\{W_k, k \geq 1\}$  sont i.i.d. de loi Gamma de paramètre 1. Donc la limite en loi de  $\sqrt{n}(Z_n/n - 1)$  est la limite en loi de

$$\sqrt{n} \left( \frac{1}{n} \sum_{k=1}^n W_k - 1 \right)$$

Puisque  $\mathbb{E}[W_1] = \text{Var}(W_1) = 1$ , on obtient le résultat demandé en appliquant le TCL pour des v.a. i.i.d.

3. ► **Solution 1 (plus rapide)** D'après la question 1,  $V_n$  suit une loi Beta de paramètres  $(k_n, n - k_n)$ . On en déduit le résultat demandé en étudiant la limite simple de sa densité (quand  $n \rightarrow \infty$ ) et en appliquant le lemme de Scheffé.

► **Solution 2** On propose une preuve basée sur l'indication suivante : établir la loi de  $X_n/Y_n$  puis observer que  $V_n = g(X_n/Y_n)$  avec  $g(x) = x/(1 + x)$

- D'après la question 1, en observant que  $X/Y = V/(1-V)$ , on obtient par la méthode d'identification

$$\mathbb{E} \left[ h \left( \frac{X}{Y} \right) \right] = \frac{\Gamma(s+t)}{\Gamma(s)\Gamma(t)} \int_0^\infty h(z) \frac{z^{s-1}}{(z+1)^{s+t}} dz$$

(on reconnaît une Loi Beta de seconde espèce). Soit  $R_n = X_n/Y_n$ ; sa densité est proportionnelle à  $z^{k_n-1}(1+z)^{-n} \mathbf{1}_{\mathbb{R}^+}(z)$ .

- Montrons que<sup>1</sup>

$$\sqrt{n} \left( R_n - \frac{p}{1-p} \right) \xrightarrow{\mathcal{L}} \mathcal{N} \left( 0, \frac{p}{(1-p)^3} \right). \quad (2)$$

Soit  $h$  une fonction continue bornée. Posons  $c = p/(1-p)$ .

$$\begin{aligned} \mathbb{E} [h(\sqrt{n}(R_n - c))] &= \frac{\Gamma(n)}{\Gamma(k_n)\Gamma(n-k_n)} \int_0^\infty h(\sqrt{n}(z-c)) z^{k_n-1} (1+z)^{-n} dz \\ &= \frac{1}{\sqrt{n}} \frac{\Gamma(n)}{\Gamma(k_n)\Gamma(n-k_n)} \int_{-c\sqrt{n}}^{+\infty} h(v) \left( \frac{v}{\sqrt{n}} + c \right)^{k_n-1} \left( 1 + c + \frac{v}{\sqrt{n}} \right)^{-n} dv \\ &= \frac{1}{\sqrt{n}} \frac{c^{k_n-1}}{(1-c)^n} \frac{\Gamma(n)}{\Gamma(k_n)\Gamma(n-k_n)} \int_{-c\sqrt{n}}^{+\infty} h(v) \left( 1 + \frac{v}{c\sqrt{n}} \right)^{k_n-1} \left( 1 + \frac{v}{(1+c)\sqrt{n}} \right)^{-n} dv \end{aligned}$$

On écrit  $\ln(1+x) = x - x^2/2 + o(x^2)$  au voisinage de zero, puis on applique le lemme de Scheffé. Pour identifier la densité limite, observer que

$$\begin{aligned} (k_n-1) \ln \left( 1 + \frac{v}{c\sqrt{n}} \right) - n \ln \left( 1 + \frac{v}{(1+c)\sqrt{n}} \right) \\ = \frac{\sqrt{n}v}{c} \left( \frac{k_n-1}{n} - \frac{c}{1+c} \right) - \frac{v^2}{2c^2} \left( \frac{k_n-1}{n} - \frac{c^2}{(1+c)^2} \right) (1+o(1)). \end{aligned}$$

Puisque  $c/(1+c) = p$ , en utilisant (1) le terme de droite converge vers (à  $v$  fixé, quand  $n \rightarrow \infty$ )

$$-\frac{v^2}{2c^2} p(1-p) = -\frac{v^2}{2} \frac{(1-p)^3}{p}.$$

On vérifie ensuite que le terme constant converge vers la constante de normalisation de  $v \mapsto \exp(-v^2(1-p)^3/(2p))$ .

- On a  $V_n = R_n/(1+R_n) = g(R_n)$  en ayant posé  $g(x) = x/(x+1)$ . Notons que  $g$  est  $C^1$  sur  $\mathbb{R}^+$  et que  $g'(x) = 1/(1+x)^2$ . En observant que

$$g \left( \frac{p}{1-p} \right) = p \quad g' \left( \frac{p}{1-p} \right) = (1-p)^2$$

on obtient le résultat demandé en appliquant la méthode delta à la convergence (2).

**Exercice 2.** Soit  $f$  une densité de probabilité portée par un intervalle (non nécessairement borné)  $(a, b) \subset \mathbb{R}$ . On suppose que  $f$  est continue et ne s'annule pas sur  $(a, b)$ . On note  $F(x) = \int_{-\infty}^x f(u)du$  la fonction de répartition associée. Cette fonction de répartition est alors strictement monotone sur  $x \in [a, b]$  et définit une bijection de  $[a, b] \rightarrow [0, 1]$ . On note  $F^{-1}$  la fonction réciproque de  $F$  de  $[0, 1] \rightarrow [a, b]$ . De plus, par continuité de  $f$ ,  $F$  est continuellement dérivable sur  $(a, b)$  de dérivée  $f$  et il s'en suit que  $F^{-1}$  est dérivable sur  $(0, 1)$ .

1. Soit  $U$  une variable uniforme sur  $[0, 1]$ ,  $U \sim \text{Unif}([0, 1])$ . Montrer que la variable  $X$  définie par  $X = F^{-1}(U)$  a pour densité  $f$ . Réciproquement, montrer que si  $X$  est une loi de densité  $f$ , alors  $U = F(X)$  est une loi uniforme sur  $[0, 1]$ .

---

1. intuiter le terme de recentrage et la variance limite pour que l'application de la méthode delta permette d'aboutir au résultat demandé.

2. Soient  $g$  une densité et  $Y_1, \dots, Y_n$ ,  $n$  v.a. i.i.d. de densité  $g$ . On note  $(Y_{(1)}, \dots, Y_{(n)})$  la statistique d'ordre de l'échantillon,  $Y_{(1)} < Y_{(2)} < \dots < Y_{(n)}$ . Montrer que  $Y_{(k)}$  a pour densité

$$g_{Y_{(k)}}(y) = \frac{n!}{(k-1)!(n-k)!} G(y)^{k-1} [1 - G(y)]^{n-k} g(y),$$

où  $G$  est la fonction de répartition associée à  $g$ . [Indication : on pourra montrer successivement  $g_{Y_{(k)}}(y) = n! \mathbb{P}(Y_1 < \dots < Y_{k-1} < y < Y_{k+1} < \dots < Y_n) g(y)$  puis  $\mathbb{P}(\max(Y_1, \dots, Y_{k-1}) < y) = (k-1)! \mathbb{P}(Y_1 < \dots < Y_{k-1} < y)$  et  $\mathbb{P}(y < \min(Y_{k+1}, \dots, Y_n)) = (n-k)! \mathbb{P}(y < Y_{k+1} < \dots < Y_n)$ .]

3. Quelle est la loi de  $Y_{(k)}$  si  $g = \mathbb{1}_{[0,1]}$  est la densité de la loi uniforme sur  $[0, 1]$  ?  
 4. Soit  $p \in (0, 1)$ . On note  $x_p$  le quantile d'ordre  $p$ , i.e.  $x_p = F^{-1}(p)$ . Montrer que

$$\sqrt{n}(X_{(k_n)} - x_p) \xrightarrow{(\text{loi})} \mathcal{N}\left(0, \frac{p(1-p)}{f^2(x_p)}\right).$$

5. Soit  $X_1, \dots, X_n$  une suite de v.a. i.i.d. normales de moyenne  $\mu$  et de variance  $\sigma^2$ . Montrer que la médiane est un estimateur consistant et asymptotiquement normal de la moyenne. Déterminer la variance asymptotique de cet estimateur. Cet estimateur doit-il être préféré à la moyenne empirique ?  
 6. Reprendre la question précédente avec  $X_1, \dots, X_n$  suite de v.a. i.i.d. distribuées suivant une loi de Laplace de densité  $f_\mu(x) = \frac{1}{2}e^{-|x-\mu|}$ . Commenter.

### Corrigé :

1. Pour tout  $x \in [a, b]$ ,  $\{F^{-1}(U) \leq x\} = \{U \leq F(x)\}$  et pour tout  $t \in [0, 1]$ , on a  $\{F(X) \leq t\} = \{X \leq F^{-1}(t)\}$ . Ce qui donne le résultat.  
 2. La correction ne suit pas l'indication : on calcule d'abord la fonction de répartition puis on obtient la densité par dérivation. On a

$$\{Y_{(k)} \leq t\} = \bigcup_{j=k}^n \{\text{il existe exactement } j \text{ variables dans } ]-\infty, t] \text{ et } (n-j) \text{ dans } ]t, \infty[ \};$$

Il s'agit d'une union disjointe et chaque événement décrit le résultat d'une Binômiale de paramètres  $(n, G(t))$ . On en déduit que

$$\mathbb{P}(Y_{(k)} \leq t) = \sum_{j=k}^n \binom{n}{j} G(t)^j (1 - G(t))^{n-j}$$

puis par dérivation, on obtient pour densité

$$t \mapsto n \binom{n-1}{k-1} (G(t))^{k-1} (1 - G(t))^{n-k} g(t)$$

3. Le cas uniforme correspond à  $g(t) = \mathbb{1}_{[0,1]}(t)$  et  $G(t) = t$  pour tout  $t \in [0, 1]$ .  
 4. Ici,  $\{k_n, n \geq 1\}$  est une suite d'entiers telle que  $\lim_n \sqrt{n}(k_n/n - p) = 0$ . On établit la propriété dans le cas où les v.a.  $\{X_k, k \geq 1\}$  sont i.i.d. uniformes sur  $[0, 1]$  (ce qui entraîne  $x_p = p$ ). Le cas général est une conséquence de la méthode delta appliquée avec la fonction  $t \mapsto F^{-1}(t)$  dont la dérivée est  $1/f(F^{-1}(t))$ .  
 D'après la question précédente,  $X_{(k_n)}$  suit une loi Beta de paramètres  $(k_n, n - k_n)$ . En utilisant l'exercice précédent, on a

$$\sqrt{n}(X_{(k_n)} - p) \xrightarrow{\mathcal{L}} \mathcal{N}(0, p(1-p)).$$



5. Par la moyenne empirique, un intervalle de confiance (non asymptotique) au niveau  $(1 - \alpha)$  est

$$\left[ n^{-1} \sum_{k=1}^n X_k \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

où  $z_{1-\alpha/2}$  est le quantile d'ordre  $1 - \alpha/2$  d'une loi  $\mathcal{N}(0, 1)$ .

Par l'estimateur de la médiane, en utilisant la question précédente, on obtient un intervalle de confiance (asymptotique) de niveau  $1 - \alpha$

$$\left[ X_{(k_n)} \pm z_{1-\alpha/2} \sqrt{\frac{\pi}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

6. D'après la question 4, il vient

$$\sqrt{n} (X_{[n/2]} - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Par le TCL, puisque  $\text{Var}(X_1) = 2$ ,

$$\sqrt{n} \left( n^{-1} \sum_{k=1}^n X_k - \mu \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 2).$$