PC 2 (Modèle statistique)

---

1 TCL pour la médiane *(Emprunté à la PC1)*

Soit $(X_1, X_2, \ldots, X_{2n+1})$ un échantillon de $2n+1$ v.a. i.i.d. uniformes sur $[0,1]$, on note Y_n la médiane de l'échantillon. On s'attend à ce que (Y_n) tende vers $1/2$, nous allons montrer que $(Y_n - 1/2)$ a des fluctuations gaussiennes.

1. Se convaincre que Y_n a pour densité
$$(2n+1)\binom{2n}{n} x^n(1-x)^n \mathbf{1}_{x\in[0,1]}.$$
2. Déterminer la densité g_n de
$$Z_n = 2\sqrt{2n}\,(Y_n - 1/2)$$
et en déduire que Z_n converge en loi vers une $\mathcal{N}(0,1)$. On rappelle la formule de Stirling : $n! \sim n^n e^{-n}\sqrt{2\pi n}$.

   (On pourra utiliser le Théorème de Scheffé : si chaque Z_n a pour densité g_n et que la suite (g_n) converge simplement vers une densité g, alors (Z_n) converge en loi vers la loi de densité g.)
3. Comment généraliser ce résultat et cette preuve à des variables continues non-uniformes ?

---

2 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 μ σ-finie sur $\mathbb{R}$ ou $\mathbb{N}$ de la forme
$$p_\theta(x) = c(\theta)\exp\left(m(\theta)f(x) + h(x)\right).$$

On supposera que Θ est un intervalle ouvert de $\mathbb{R}$, $m(\theta)=\theta$ et $c(\cdot)\in C^2$, $c(\theta)>0$ pour tout $\theta\in\Theta$. On notera X une variable aléatoire de loi $\mathbb{P}_\theta$ et on admettra que
$$\frac{d^i}{d\theta^i}\int \exp(\theta f(x)+h(x))\,\mu(dx) = \int f(x)^i \exp(\theta f(x)+h(x))\,\mu(dx) < +\infty, \quad \text{pour } i=1,2.$$

1. Montrez que $\varphi(\theta) := \mathbb{E}_\theta\left(f(X)\right) = -\frac{d}{d\theta}\log(c(\theta))$.
2. Montrez que $\mathrm{Var}_\theta(f(X)) = \varphi'(\theta) = -\frac{d^2}{d\theta^2}\log(c(\theta))$.
3. On dispose d'un n-échantillon $X_1,\ldots,X_n$ 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 $\mathrm{Var}_\theta(f(X))>0$, montrez que
$$\sqrt{n}\left(\hat\theta_n - \theta\right) \xrightarrow{\text{loi}} \mathcal{N}\left(0, \frac{1}{\mathrm{Var}_\theta\left(f(X)\right)}\right).$$

3 Estimation par la méthode plug-in

Soit $X_1, \ldots, X_n$ 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 $\widehat{\theta}$ de θ.
2. Déterminer l'estimateur plug-in de la variance de $\widehat{\theta}$ et en déduire un intervalle de confiance asymptotique pour θ de niveau $1 - \alpha$.

4 Stabilisation de la variance

On dispose d'un échantillon $X_1, \ldots, X_n$ 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_α 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 $\lim_{n\to\infty} \mathbb{P}(\theta \in \hat{I}_{n,\alpha}) = 1 - \alpha$.

5 Modèle d'autorégression

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

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

avec les ξ_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.

6 Survie

Un système fonctionne en utilisant deux machines de types différents. Les durées de vie X_1 et X_2 des deux machines suivent des lois exponentielles de paramètres λ_1 et λ_2. Les variables aléatoires X_1 et X_2 sont supposées indépendantes.

1. Montrer que une variable aléatoire X suit la loi exponentielle $\mathcal{E}(\lambda)$ si et seulement si

$$\forall x > 0 : \ \mathbb{P}(X > x) = \exp(-\lambda x).$$

2. 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.
3. 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.
4. On dispose de n systèmes identiques et fonctionnant indépendamment les uns des autres dont on observe les durées de vie $Z_1, \ldots, Z_n$.
   (a) Écrire le modèle statistique correspondant. Les paramètres λ_1 et λ_2 sont-ils identifiables ?
   (b) Supposons maintenant que l'on observe à la fois les durées de vie des systèmes $Z_1, \ldots, Z_n$ et les causes de la défaillance correspondantes $I_1, \ldots, I_n$, $I_i \in \{0, 1\}$. Écrire le modèle statistique dans ce cas. Les paramètres λ_1 et λ_2 sont-ils identifiables ?

7 Problème bonus : Quantile central

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

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

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

   $$\sqrt{n} \left(\frac{Z_n}{n} - 1 \right) \Rightarrow_d \mathcal{N}(0, 1) \;.$$

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

   $$\sqrt{n} \left(\frac{k_n}{n} - p \right) \to 0 \;. \tag{1}$$

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

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

   Montrer que

   $$\sqrt{n} \left(V_n - p \right) \Rightarrow_d \mathcal{N}(0, p(1-p)) \;.$$

   [*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.

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'annulle 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 $[a, b] \to [0, 1]$. On note F^{-1} la fonction réciproque de F de $[0, 1] \to [a, b]$. De plus, par continuité de f, F est continuement 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 $[0, 1]$, $U \sim$ 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 $[0, 1]$.

2. Soient g une densité et $Y_1, \ldots, Y_n$, n v.a. i.i.d. de densité g. On note $(Y_{(1)}, \ldots, Y_{(n)})$ la statistique d'ordre de l'échantillon, $Y_{(1)} < Y_{(2)} < \cdots < Y_{(n)}$. 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 $g_{Y_{(k)}}(y) = n! \mathbb{P}(Y_1 < \cdots < Y_{k-1} < y < Y_{k+1} < \cdots < Y_n) g(y)$ puis $\mathbb{P}(\max(Y_1, \ldots, Y_{k-1}) < y) = (k-1)!\ \mathbb{P}(Y_1 < \cdots < Y_{k-1} < y)$ et $\mathbb{P}(y < \min(Y_{k+1}, \ldots, Y_n)) = (n-k)!\ \mathbb{P}(y < Y_{k+1} < \cdots < Y_n)$.]
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
$$\sqrt{n}(X_{(k_n)} - x_p) \Rightarrow_d \mathcal{N}\left(0, \frac{p(1-p)}{f^2(x_p)}\right) \ .$$
5. Soit $X_1, \ldots, X_n$ une suite de v.a. i.i.d. normales de moyenne μ et de variance σ^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 $X_1, \ldots, X_n$ suite de v.a. i.i.d. distribuées suivant une loi de Laplace de densité $f_\mu(x) = \frac{1}{2} \mathrm{e}^{-|x-\mu|}$. Commenter.