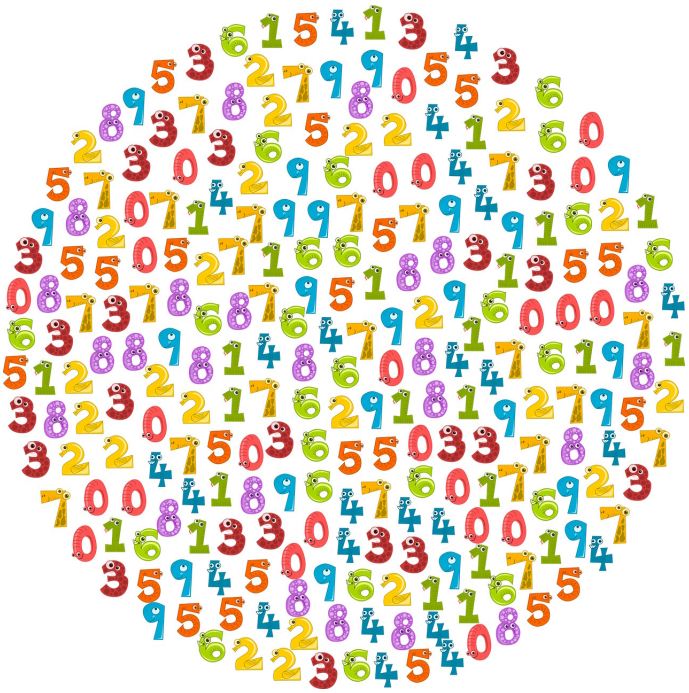
Projets de simulation Python

MAP361P - *Aléatoire*



2023-2024

PROJET 9

Détection de variables influentes, construction d'une règle de décision

sujet proposé par Thierry Klein

thierry.klein@math.univ-toulouse.fr

Les gros codes numériques de simulations sont de plus en plus utilisés par les industriels pour simuler des phénomènes complexes. Ces codes sont complexes et se comportent comme des boîtes noires. Mathématiquement ils sont vus comme une fonction f inconnue qui à un jeu d'entrées $(x_1, \ldots, x_d)$ renvoie une sortie $y = f(x_1, \ldots, x_d)$. Une des problématiques des industriels est de comprendre quels sont parmi les variables $(x_1, \ldots, x_d)$ celles qui influent le plus sur y. Pour faire cela, ils quantifient l'importance des variables d'entrées à l'aide d'indices. L'un des indices le plus utilisé est l'indice de Sobol dont nous allons étudier certaines propriétés dans ce projet.

1 Préliminaires: méthode de construction d'une règle de décision

Soit θ un paramètre réel inconnu et on cherche à savoir si

$$H_0 : \text{“}\theta = \theta_0\text{”}$$

que nous appelerons l'hypothèse H_0 ou si

$$H_1 : \text{“}\theta < \theta_0\text{”}$$

que nous appelerons l'hypothèse alternative (le cas $\theta > \theta_0$ se traite de façon analogue). Supposons que nous disposions d'un estimateur $\widehat{\theta}_n := \widehat{\theta}(X_1, \ldots, X_n)$ fonction de n variables aléatoires iid qui convergent

presque sûrement vers θ. Une règle de décision est une fonction T qui à $(X_1, \dots, X_n)$ renvoie 0 ou 1 (si $T = 0$ on décidera H_0 et si $T = 1$ on décidera H_1). Soit K un réel et $T_K = 1_{\{\widehat{\theta}_n < K\}}$. Soit $\alpha \in]0, 1[$, l'idée est de trouver K tel que

$$\mathbb{P}_{\theta=\theta_0}(T_K = 1) \leqslant \alpha. \tag{9.1}$$

C'est-à-dire que la probabilité de décider H_1 alors que H_0 est vraie est plus petite qu'un seuil α (la notation $\mathbb{P}_{\theta=\theta_0}$ signifiant que l'on se place sous l'hypothèse que H_0 est vraie).

T1. Soit $X_1, \dots, X_n$ des variables aléatoires iid de loi $\mathcal{N}(\theta, 1)$ où θ est inconnu.

1. (a) Montrer que $\widehat{\theta}_n = \frac{1}{n}\sum_{k=1}^n X_k$ converge presque sûrement vers θ.
   (b) Déterminer la loi de $\sqrt{n}(\widehat{\theta}_n - \theta)$.
2. Soit $\alpha \in]0, 1[$ fixé et Φ la fonction répartition de la loi Gaussienne $\mathcal{N}(0, 1)$. Déterminer en fonction de α, de n, de Φ^{-1} et de θ_0 la valeur optimale de K_α pour que la relation (9.1) soit satisfaite avec une égalité.

S1. Créer un code python permettant:

1. de simuler n variables gaussiennes indépendantes de loi $\mathcal{N}(\theta_0, 1)$,
2. de calculer pour α et θ_0 fixés, la constante K_α et qui vous renvoie la valeur optimale de votre règle de décision.
3. Pour $\theta_0 = 3,\ \alpha = 0.05$ et $n = 100$ répéter $N = 100$ fois l'expérience et compter le nombre fois où vous avez décidé H_0. On appelle ce nombre N_n quelle est sa loi?
4. Pour $n \in \{100, 1000\}$ et $N \in \{100, 1000, 5000\}$, Simuler 500 valeurs de la variable aléatoire N_n et les représenter sous forme d'histogramme. Que remarquez-vous?

2 Rappel sur l'espérance conditionnelle

Dans tout ce projet on supposera que toutes les variables sont à densités par rapport à la mesure de Lebesgue. Nous rappelons les résultats suivants sur l'espérance conditionnelle.

Théorème 2.1 (**et Définition**). *Soit Y une variable aléatoire telle que $\mathbb{E}[Y^2] < \infty$ et X un vecteur aléatoire. Alors il existe p.s. une unique v.a. X-mesurable[1] notée $\mathbb{E}[Y|X]$ vérifiant pour toute fonction h mesurable bornée (ou dans L^2)*

$$\mathbb{E}\left[h(X)\mathbb{E}(Y|X)\right] = \mathbb{E}[h(X)Y].$$

On appelle la v.a. $\mathbb{E}[Y|X]$ l'espérance conditionnelle de Y sachant X.
Un résultat important de Doob assure que $\mathbb{E}[Y|X]$ peut s'écrire comme une fonction mesurable de X: c'est-à-dire qu'il existe une fonction f telle que

$$\mathbb{E}[Y|X] = f(X).$$

Proposition 2.1. *L'espérance conditionnelle est linéaire, De plus*

1. *Si $Z = g(X)$ alors*
$$\mathbb{E}\left[ZY|X\right] = Z\mathbb{E}\left[Y|X\right].$$
2. *Si Y est une variable aléatoire indépendante de X*
$$\mathbb{E}\left[Y|X\right] = \mathbb{E}\left[Y\right].$$
3. *Si X est un vecteur aléatoire et X' est un sous vecteur de X (on extrait de X un certain nombre de coordonnées) alors*
$$\mathbb{E}\left[Y|X'\right] = \mathbb{E}\left[\mathbb{E}\left[Y|X\right]|X'\right].$$
En particulier $\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$.

[1] C'est-à-dire mesurable par rapport à $\sigma(X)$ la plus petite tribu qui rend X-mesurable.

T2. Soit (X_1, X_2, X_3) trois variables aléatoires indépendantes, calculer pour tout $(i,j) \in \{1,2,3\}^2$ $\mathbb{E}[Y|X_i]$ et $\mathbb{E}[Y|X_i, X_j]$ lorsque

1. $Y = f(X_1, X_2, X_3) = aX_1 + X_2 + X_3 + bX_2X_3$ où a et b sont deux paramètres réels et les X_i sont de lois gaussiennes $\mathcal{N}(0,1)$.
2. $Y = g(X_1, X_2, X_3) = \sin X_1 + 7\sin^2 X_2 + 0.1X_3^4 \sin X_1$ où les X_i sont de lois uniformes sur $[-\pi, \pi]$
3. $Y = h(X_1, X_2, X_3) = \exp(X_1 + 2X_2 + aX_3)$ où a est un paramètre réel et les X_i sont de lois gaussiennes $\mathcal{N}(0,1)$.

3 Indices de Sobol et présentation du problème

Soit f une fonction de $\mathbb{R}^p$ dans $\mathbb{R}$ et $X_1, \ldots, X_p$ des variables aléatoires indépendantes n'ayant pas forcément la même loi. On pose

$$Y = f(X_1, \ldots, X_p).$$

Pour $u \subset \{1, 2, \ldots, p\}$, on définit l'indice de Sobol S_u par

$$S_u = \frac{\mathrm{Var}(\mathbb{E}[Y|X_i, i \in u])}{\mathrm{Var}(Y)}. \tag{9.2}$$

Par convention si $u = \emptyset$ alors $\mathbb{E}[Y|X_i, i \in u] = \mathbb{E}[Y]$. Cette quantité représente la part de la variance de Y qui est dûe à l'action des variables X_i pour $i \in u$. Les variables $X_i, i \in u$ seront d'autant plus influentes que S_u sera proche de 1.

T3. Dans la suite du projet si $w \subset \{1, 2, \ldots, p\}$, l'ensemble des $\{X_i,\ i \in w\}$ sera noté par X_w.

1. Soit $w \subset \{1, 2, \ldots, p\}$ et notons $\sim w$ le complémentaire de w dans $\{1, 2, \ldots, p\}$ en remarquant que

$$\begin{aligned} Y =& \mathbb{E}[Y] + (\mathbb{E}[Y|X_w] - \mathbb{E}[Y]) + (\mathbb{E}[Y|X_{\sim w}] - \mathbb{E}[Y]) \\ &+ Y - (\mathbb{E}[Y] + (\mathbb{E}[Y|X_w] - \mathbb{E}[Y]) + (\mathbb{E}[Y|X_{\sim w}] - \mathbb{E}[Y]))\,., \end{aligned} \tag{9.3}$$

   montrer que

$$\mathrm{Var}(Y) = \mathrm{Var}(\mathbb{E}[Y|X_w]) + \mathrm{Var}(\mathbb{E}[Y|X_{\sim w}]) + \mathrm{Var}(Y - \mathbb{E}[Y|X_w] - \mathbb{E}[Y|X_{\sim w}]). \tag{9.4}$$

2. En déduire que $0 \leqslant S_w \leqslant 1$,
3. Soit maintenant $v \subset u \subset \{1, 2, \ldots, p\}$. D'après le résultat de Doob on sait qu'il existe une fonction g telle que $\mathbb{E}[Y|X_u] = g(X_u)$. En appliquant (9.3) et (9.4) à la fonction g en remplaçant $\{1, \ldots, p\}$ par u et w par v montrer que $S_v \leqslant S_u$.

Dans beaucoup de situations réelles la fonction f n'est pas connue explicitement et le calcul exacte des valeurs de S_u n'est pas possible. L'objectif de la suite de ce projet est de contruire une règle de décision pour le problème de décision

$$H_0 : \text{“}S_v = S_u\text{”}$$

contre l'alternative

$$H_1 : \text{“}S_v < S_u\text{”}.$$

Principe de construction de la règle

- Etape 1: Construire à partir d'un échantillon de taille n des estimateurs $\widehat{S}_v$ et $\widehat{S}_u$ de S_v et de S_v.
- Etape 2: Etablir que $\sqrt{n}\left((\widehat{S}_v, \widehat{S}_u) - (S_v, S_u)\right)$ converge en loi vers un vecteur gaussien de dimension 2 dont on déterminera la matrice de covariance Γ.
- Etape 3: Déterminer la valeur optimale de K pour laquelle la règle de décision $T_K = 1_{\{\widehat{S}_v - \widehat{S}_u < K\}}$ vérifie la propriété (9.1).

4 Estimation des indices S_u

Soit $(X_1, X_2, \ldots, X_p)$ des variables aléatoires indépendantes et $(X'_1, X'_2, \ldots, X'_p)$ une copie indépendantes de $(X_1, X_2, \ldots, X_p)$ (cela signifie que les variables X'_i sont indépendantes entre elles, indépendantes des variables X_i et que X'_i et X_i ont la même loi). Soit $Y = f(X_1, X_2, \ldots, X_p)$ telle que $\mathrm{Var}(Y)$ existe.

T4. Montrer que

$$\mathrm{Var}(\mathbb{E}[Y|X_i, i \in u]) = \mathrm{Cov}(f(X_1, X_2, \ldots, X_p), f(X_1^u, X_2^u, \ldots, X_p^u))$$

où $X_i^u = X_i$ si $i \in u$ et $X_i^u = X'_i$ sinon.
Soit maintenant deux sous ensembles $v \subset u \subset \{1, 2, \ldots, p\}$ et soit $(X_1(k), X_2(k), \ldots, X_p(k))_{1 \leqslant k \leqslant n}$ et $(X'_1(k), X'_2(k), \ldots, X'_p(k))_{1 \leqslant k \leqslant n}$ deux n-échantillons de $(X_1, X_2, \ldots, X_p)$ et de $(X'_1, X'_2, \ldots, X'_p)$. On pose pour tout $1 \leqslant k \leqslant n$

$$\begin{aligned} Y_k &= f(X_1(k), X_2(k), \ldots, X_3(k)), \\ Y_k^{(u)} &= f(X_1^u(k), X_2^u(k), \ldots, X_p^u(k)), \\ Y_k^{(v)} &= f(X_1^v(k), X_2^v(k), \ldots, X_3^v(k)). \end{aligned}$$

Et on considère

$$\begin{pmatrix} S_n^u \\ \\ S_n^v \end{pmatrix} = \begin{pmatrix} \frac{\frac{1}{n}\sum_{k=1}^n Y_k Y_k^{(u)} - \frac{1}{n}\sum_{k=1}^n Y_k \frac{1}{n}\sum_{k=1}^n Y_k^{(u)}}{\frac{1}{n}\sum_{k=1}^n Y_k^2 - (\frac{1}{n}\sum_{k=1}^n Y_k)^2} \\ \\ \frac{\frac{1}{n}\sum_{k=1}^n Y_k Y_k^{(v)} - \frac{1}{n}\sum_{k=1}^n Y_k \frac{1}{n}\sum_{k=1}^n Y_k^{(v)}}{\frac{1}{n}\sum_{k=1}^n Y_k^2 - (\frac{1}{n}\sum_{k=1}^n Y_k)^2} \end{pmatrix}. \tag{9.5}$$

T5. Loi forte des grands nombres : Montrer que les estimateurs sont consistants i.e. le vecteur (S_n^u, S_n^v) converge presque sûrement vers (S_u, S_v).

T6. On admettra cette version de la delta méthode

Théorème 4.1 (Méthode Delta)**.** *Soit ϕ une application de $\mathbb{R}^k$ dans $\mathbb{R}^m$ différentiable en θ. Soit T_n des vecteurs aléatoires de $\mathbb{R}^k$ (à valeurs dans le domaine de définition de ϕ) et $(r_n)_n$ une suite de nombres réels tendant vers ∞. Alors*

$$r_n\left(\phi(T_n) - \phi(\theta)\right) \xrightarrow[n\to+\infty]{\mathcal{L}} D\phi(\theta)(T),$$

dès que $r_n\left(T_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} T$.

Soit $(Z_k)_{k \geqslant 1}$ des vecteur aléatoires i.i.d. de $\mathbb{R}^k$ de moyenne θ et de matrice de covariance Σ. Appliquer le Théorème central limite vectoriel aux vecteurs Z_k et en déduire la loi limite de

$$\sqrt{n}\left(\phi(\frac{1}{n}\sum_{k=1}^n Z_k) - \phi(\theta)\right).$$

T7. Théorème central limite: L'objectif de cette question est de ontrer que si $\mathbb{E}[Y^4]$ existe alors les estimateurs sont asymptotiquement normaux i.e. $\sqrt{n}((S_n^u, S_n^v) - (S_u, S_v))$ converge en loi vers un vecteur gaussien centré de dimension 2. Puis d'expliciter la matrice de covariance Γ.

1. Appliquer le théorème central limite vectoriel au vecteur aléatoire $\mathbb{R}^6$

$$Z_k = (Y_k, Y_k^{(u)}, Y_k^{(v)}, Y_k Y_k^{(u)}, Y_k Y_k^{(v)}, Y_k^2)$$

et déterminer le vecteur m et la matrice Σ telle que

$$\sqrt{n}\left(\frac{1}{n}\sum_{k=1}^{n} Z_k - m\right) \xrightarrow{\mathcal{L}} \mathcal{N}_6(m, \Sigma).$$

2. Conclure en appliquant la delta méthode à la fonction ϕ de $\mathbb{R}^6$ dans $\mathbb{R}^2$ définie par

$$\phi(x_1, x_2, \ldots, x_6) = \left(\frac{x_4 - x_1 x_2}{x_6 - x_1^2}, \frac{x_5 - x_1 x_3}{x_6 - x_1^2}\right).$$

3. Exprimer Γ en fonction de Σ et de la matrice Jacobienne de ϕ.

S2. Créer un code python permettant de calculer les estimateurs (S_n^u, S_n^u). Puis faire calculer pour $n = 1000$ les valeurs de (S_n^u, S_n^v) pour la fonction

$$f(X_1, X_2, X_3, X_4) = \sum_{i=1}^{4} X_i + bX_1X_2 + cX_1X_2X_3$$

pour $u = \{1, 2, 3\}, v = \{1, 2\}$ (on prendra dans un premier temps $b = 2$ et $c = 3$ puis $b = 2$ et $c = 0$).

S3. Créer un code python permettant de tracer les 2 courbes $n \mapsto S_n^u$, $n \mapsto S_n^v$, pour n variant de 10 à 10^4. On affichera pour la fonction f de la question précédente les 2 courbes. Ces courbes sont-elles cohérentes par rapport aux résultats de la question T5?

S4. Créer un code python permettant d'estimer la matrice de covariance Γ. Calculer cet estimateur sur la même fonction. Dans la suite on notera $\widehat{\Gamma}$ cet estimateur.

T8. Montrer lorsque H_0 est vraie que la loi limite de $\sqrt{n}(S_n^u - S_n^v)$ est une gaussienne centrée. Exprimer la variance σ^2 de cette gaussienne en fonction de la matrice Γ.

S5. Créer un code python permettant de calculer $N = 10^4$ valeur distinctes de la quantité $\sqrt{n}(S_n^u - S_n^v)$ pour $n = 10^4$. Tracer les histogrammes de ces N valeurs pour la fonction précédentes pour les 2 choix de b et c. Que remarque t-on?

S6. Créer un code python permettant de calculer un estimateur $\widehat{\sigma^2}$ consistant de σ^2 et reprendre la question précédente en remplaçant $\sqrt{n}(S_n^u - S_n^v)$ par $\frac{\sqrt{n}(S_n^u - S_n^v)}{\sqrt{\widehat{\sigma^2}}}$. Que remarque t on?

T9. Déterminer la valeur optimale de K pour laquelle la règle de décision $T_K = 1_{\{\widehat{S}_v - \widehat{S}_u < K\}}$ vérifie la propriété (9.1). On prendra $\alpha = 0.05$.

S7. Créer un code python permettant pour $n = 500$ et pour les 2 choix de b et c de la question S2

1. de calculer $N = 1000$ valeurs de T_K,
2. de compter le nombre de fois où $T_K = 1$, on notera $N_K(n)$ ce nombre.

Ces résultats sont ils cohérents avec la théorie?

S8. Reprendre le code de la question précédente en faisant varier n de 100 à 5000 et tracer en fonction de n les 2 courbes $n \mapsto N_K(n)$. Que remarquez vous?