Université Rennes 1
Master 1 Mathématiques année 2022-2023
Nathalie Krell

Statistiques

Fiche de TD n°1 :

$\boxed{1}$ **Simulation.** On suppose que l'on sait simuler des nombres aléatoires uniformément répartis sur $[0,1]$. Donner une méthode de tirage de nombres aléatoires répartis sur $[0,1]$. Donner une méthode de tirage de nombres aléatoires suivant une loi de probabilité ν donné.

[a] Si ν est une loi discrète.

[b] Si ν est une loi sur $\mathbb{R}$ de f.r. F.
Indication On pourra définir une application G de telle sorte que $G(t) < x \Leftrightarrow t < F(x)$.

$\boxed{2}$ **Loi de $F(X)$.**
Soit X une v.a.r. de f.r. F. On suppose F continue. Donner la loi de probabilité de $F(X)$. (On pourra commencer par supposer que F est strictement croissante).

$\boxed{3}$ Néo décide de tricher à ses examens. On suppose qu'à chaque fois, ses triches sont indépendantes. A chaque fois que Néo triche, il a une probabilité p de réussir.
Soit T le nombre de fois où Néo a réussi à tricher avant d'échouer pour la première fois.
1) Quelle est la loi de T ?
2) Ecrivez le modèle sous la forme d'un triplet, espace d'observation, tribu, famille de probabilités.
3) Le modèle est-il identifiable ? Dominé ?

$\boxed{4}$ Un circuit électrique est composé de deux types de diodes A et B montées en série. Les durées de vie des diodes, qui sont indépendantes, suivent des lois exponentielles de paramètres inconnus éventuellement différents.

[a] Quelle est la loi suivie par la durée de vie du circuit ?

[b] On a observé les durées de vie de n circuits indépendants de ce type. Quel est le modèle statistique associé à cette expérience ?

[c] Dans l'expérience de la question précédente, on a seulement observé le type de diode (A ou B) qui a défailli. Déterminer le modèle statistique associé à cette expérience.

$\boxed{5}$ On considère deux variables aléatoires réelles indépendantes X et R définies sur l'espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$, telles que R est une variable aléatoire de Rademacher, i.e. $\mathbb{P}(R=1) = \mathbb{P}(R=-1) = 1/2$, et X suit la loi $\mathcal{N}(0,1)$. Pour $\theta \in \mathbb{R}$, on note $Y_\theta = \theta R + X$.

[a] Calculer la loi de Y_θ.

[b] Soit une expérience dans laquelle on observe n réalisations indépendantes de la loi de Y_θ , avec θ inconnu. Préciser le modèle statistique. En donner une mesure dominante et la famille des densités associées.

$\boxed{6}$ D'une génération à la suivante, le poids moyen des cochons d'un élevage augmente d'un facteur multiplicatif inconnu, mais cette tendance est perturbée de façon additive par une variable aléatoire réelle gaussienne centrée, indépendante des générations précédentes et de variance inconnue indépendante de la génération. Quel est le modèle statistique associé à l'observation du poids moyen des cochons de l'élevage sur n générations ?

$\boxed{7}$ Soit le modèle statistique d'échantillonnage,

$$X_1, \ldots, X_n \text{ i.i.d. } \sim \mathcal{N}(m, \sigma^2),\ \theta = (m, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*.$$

Parmi les variables aléatoires suivante, lesquelles sont des statistiques ?
— la moyenne empirique $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$,
— $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$
— $n\hat{\sigma}^2/\sigma^2$,
— $T = \sqrt{n-1}\left(\bar{X} - m\right)/\hat{\sigma}$.

$\boxed{8}$ Décrire l'expérience statistique $\mathcal{E}$, étudier l'identifiabilité et la domination du modèle dans chacun des cas suivants :

[a] X suit une loi $\mathcal{U}[0, \theta], \theta \in \mathbb{R}_+^*$.

[b] X suit une loi uniforme $\{0, 1, ..., \theta\}$, $\theta \in \mathbb{N}^*$.

[c] X suit une loi $\mathcal{N}(\nu, \sigma^2)$, $\theta = (\nu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$.

Soit l'expérience produit $\mathcal{E}^{\otimes n}$ obtenue en considérant n copies indépendantes de $\mathcal{E}$. Etudier alors dans chacun des cas précédents l'indentifiabilité et la domination du modèle.

$\boxed{9}$ On observe $X_1, X_2, ..., X_n$ n v.a. i.i.d. de loi $\mathcal{N}(\nu, \sigma^2)$, σ^2 connue et ν paramètre réel inconnu.

[a] Décrire l'expérience statistique $\mathcal{E}_1$.

[b] Écrire le modèle sous forme d'un modèle linéaire, c'est à dire $X = M\nu + \gamma\epsilon$ où $\epsilon \sim \mathcal{N}(0, Id_n)$.

[c] Proposer un estimateur de ν et étudier ses propriétés asymptotiques (limite en probabilité, dans $\mathbb{L}^2$,...)

[d] On dispose de l'information supplémentaire suivante : on observe également $Y = \mathbb{1}_{\{\nu \geq 0\}}$. Écrire l'expérience statistique correspondante, $\mathcal{E}_2$. Écrire l'expérience $\mathcal{E} = \mathcal{E}_1 \otimes \mathcal{E}_2$. Proposer un nouvel estimateur de ν dans l'expérience $\mathcal{E}$.

$\boxed{10}$ Soit $\epsilon_1, \epsilon_2, ..., \epsilon_n$ n v.a. i.i.d. de loi $\mathcal{N}(0, 1)$. On suppose qu'on observe

$$\begin{aligned} Y_1 &= \nu_1 + \sigma\epsilon_1 \\ Y_2 &= \nu_2 + \frac{\sigma}{\sqrt{3}}\epsilon_2 \\ Y_3 &= \nu_3 + \sigma\epsilon_3 \\ Y_4 &= \nu_4 + \frac{\sigma}{\sqrt{2}}\epsilon_4 \end{aligned}$$

Écrire la représentation linéaire du modèle c'est à dire $Y = M\theta + \gamma\epsilon$ où $\epsilon \sim \mathcal{N}(0, Id_4)$.

$\boxed{11}$ Soient $X_1, X_2, ..., X_n$ n v.a. i.i.d. de loi $\mathbb{P}_\theta(dx) = e^{\theta - x}\mathbb{1}_{\{[\theta, \infty[\}}(x)dx$, $\theta \in \mathbb{R}_+^*$.

[a] Écrire le modèle. Est-il identifiable ? dominé ?

[b] Calculer $\mathbb{E}_\theta(X_1)$ et en déduire un estimateur de θ que l'on notera $\hat{\theta}_n$.

[c] Étudier $\mathbb{E}_\theta((\hat{\theta}_n - \theta)^2)$.

[d] Trouver un meilleur estimateur (et prouver qu'il est meilleur, dans un sens que l'on cherchera à préciser).

$\boxed{12}$ **Statistiques d'ordre.*** Sur un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$, on considère n v.a.r. $X_1, X_2, ..., X_n$ indépendantes de même loi de densité f sur $\mathbb{R}$, de fonction de répartition F : le n-uplet $(X_1, ..., X_n)$ est appelé "n-échantillon" de la loi de densité f. Si $\omega \in \Omega$ est tel que les nombres réels $X_1(\omega), ..., X_n(\omega)$ sont deux à deux distincts, on définit $\sigma(\omega)$ comme la permutation de l'ensemble $\{1, 2, ..., n\}$ telle que $X_{\sigma(\omega)(1)} < X_{\sigma(\omega)(2)} < ...X_{\sigma(\omega)(n)}$.

I (a) Montrer que l'application $\omega \mapsto \sigma(\omega)$ est définie $\mathbb{P}$.p.s. sur Ω, est mesurable, et donner sa loi de probabilité.

(b) Calculer, pour $1 \leq i \leq n$ et $1 \leq k \leq n$, $\mathbb{P}(\sigma(i) = k)$.

II On note pour simplifier $X_{\sigma(\omega)(i)}(\omega) = X_{(i)}(\omega)$, pour $i = 1, ..., n$.

(a) Vérifier que $X_{(1)}(\omega), ..., X_{(n)}(\omega)$ sont des v.a. définies $\mathbb{P}$-p.s. sur Ω et que le n-uplet $(X_{(1)}(\omega), ..., X_{(n)}(\omega))$ admet une densité que l'on calculera en fonction de f. Ce n-uplet est appelé "n-échantillon réordonné" de la loi de densité f.

(b) Calculer la loi de probabilité de $X_{(1)}$, de $X_{(n)}$, du couple $(X_{(1)}, X_{(n)})$ et enfin de $X_{(i)}$, $2 \leq i \leq n-1$, en fonction de f et F.

$\boxed{13}$* Soient $(P_n)_{n\geq 1}$ et P des lois de probabilités sur $\mathbb{R}^d$ telles que pour chaque $n \geq 1$, P_n et P possèdent des densités par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$, notées respectivement f_n et f. On s'intéresse à la distance en variation totale entre ces deux probabilités :

$$d_V(P_n, P) = \sup_{A \in \mathcal{B}(\mathbb{R}^d)} |P_n(A) - P(A)|$$

Les deux premières questions sont habituellement identifiées sous le nom de théorème de Scheffé.

[a] Établir la relation

$$d_V(P_n, P) = \frac{1}{2} \int_{\mathbb{R}^d} |f_n(x) - f(x)| dx.$$

[b] En déduire que $(P_n)_{n\geq 1}$ converge vers P en variation totale si la suite $(f_n)_{n\geq 1}$ converge presque partout vers f .

[c] Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires réelles indépendantes définies sur l'espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ telle que pour chaque $n \geq 1$, X_n suit la loi $\mathcal{N}(m, \sigma^2)$, avec $m \in \mathbb{R}$ et $s > 0$. On note

$$\widehat{m_n} = \frac{1}{n} \sum_{i=1}^{n} X_i \text{ et } \widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \widehat{m_n})^2$$

Montrer que $\mathbb{P}$-p.s., la suite de lois $\mathcal{N}(\widehat{m_n}, \widehat{\sigma^2})$ converge en variation totale vers $\mathcal{N}(m, \sigma^2)$.