JUNIA$^{\text{ISEN}}$ $*$ 2 avril 2021

Probabilités & statistiques

Consignes :

- Vous disposez de **2 h** pour répondre aux **4** $\times$ **3** questions suivantes.
- **Calculatrice** non programmable **autorisée**, voire conseillée.
- Un **aide-mémoire** comportant certains faits à (ne pas) connaître est fourni en annexe.
- Soyez **précis** et **détaillé** dans vos réponses et justifications ; **exprimez**-vous !

Exercice 1

Considérons un couple aléatoire (X, Y) uniformément distribué dans le disque $\mathcal{D} : x^2 + y^2 \leqslant 1$.

a) Montrer que X et Y sont des variables aléatoires continues de densité

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

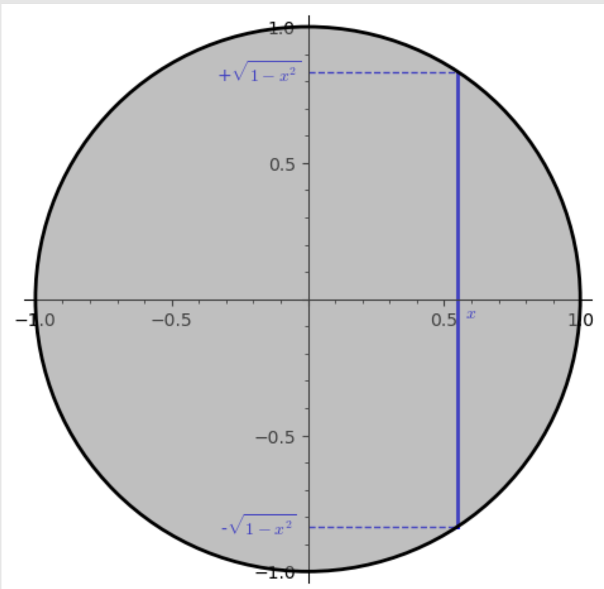
Le couple (X, Y) est uniformément distribué dans $\mathcal{D}$, sa densité de probabilité conjointe est donnée par

$$f(x,y) = \begin{cases} \frac{1}{\text{aire}(\mathcal{D})} = \frac{1}{\pi} & \text{si } (x,y) \in \mathcal{D}, \\ 0 & \text{sinon.} \end{cases}$$

On obtient donc les densités marginales par intégration :

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}y = \int_{-\sqrt{1-x^2}}^{+\sqrt{1-x^2}} \frac{1}{\pi}\,\mathrm{d}y = \frac{2}{\pi}\sqrt{1-x^2} \qquad \text{pour } x \in [-1,1],$$

et similairement pour Y. Cette densité est connue sous le nom de loi du demi-cercle.



b) Déterminer $\mathbb{E}[X]$ et $\mathrm{Var}(X)$.

[*Indication :* le changement de variables $x = \cos\theta$ peut être utile dans les intégrales.]

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f_X(x)\,\mathrm{d}x = \frac{2}{\pi}\int_{-1}^{1} x\sqrt{1-x^2}\,\mathrm{d}x = 0 \qquad \text{par symétrie}$$

$$\mathbb{E}[X^2] = \int_{-\infty}^{+\infty} x^2 f_X(x)\,\mathrm{d}x = \frac{4}{\pi}\int_{0}^{1} x^2\sqrt{1-x^2}\,\mathrm{d}x = \frac{4}{\pi}\int_0^{\frac{\pi}{2}} \cos^2\theta\sin^2\theta\,\mathrm{d}\theta$$

$$= \frac{1}{\pi}\int_0^{\frac{\pi}{2}} \sin^2(2\theta)\,\mathrm{d}\theta = \frac{1}{\pi}\int_0^{\frac{\pi}{2}} \frac{1-\cos(4\theta)}{2}\,\mathrm{d}\theta = \frac{1}{4}$$

d'où $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1}{4}$.

c) Les variables aléatoires X et Y sont-elles corrélées ? Indépendantes ?

Par symétrie (la fonction $(x,y) \mapsto xyf(x,y)$ étant impaire par rapport aux axes de coordonnées) on a $\mathbb{E}[XY] = 0$ et donc $\mathrm{Cor}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0$: les variables ne sont pas corrélées.

Par contre elles ne sont pas indépendantes, puisque $f(x,y) \neq f_X(x)\,f_Y(y)$.

Exercice 2

a) Une variable aléatoire X suit une loi dite *log-normale* de paramètres $\mu \in \mathbb{R}$ et $\sigma > 0$ lorsque $Z := \ln X$ suit une loi $\mathcal{N}(\mu, \sigma^2)$. Exprimer les fonctions F_X puis f_X en termes de F_Z et f_Z.

$$F_X(x) = \mathbb{P}[X \leqslant x] = \mathbb{P}[e^Z \leqslant x] = \mathbb{P}[Z \leqslant \ln x] = F_Z(\ln x)$$

d'où

$$f_X(x) = \frac{\mathrm{d}F_X}{\mathrm{d}x} = \frac{1}{x}\cdot f_Z(\ln x) \qquad (x > 0).$$

b) Montrer que X comme en a) est de médiane e^μ et d'espérance $e^{\mu+\frac{\sigma^2}{2}}$.

Médiane : c'est la valeur de x pour laquelle $F_X(x) = F_Z(\ln x) = \frac{1}{2}$. Puisque la médiane de Z est son espérance μ, on trouve $\ln x = \mu$ donc $x = e^\mu$.

Espérance : on calcule

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f_X(x)\,\mathrm{d}x = \int_0^{\infty} f_Z(\ln x)\,\mathrm{d}x = \int_{-\infty}^{+\infty} f_Z(z)\,e^z\,\mathrm{d}z = e^{\mu+\frac{\sigma^2}{2}}$$

en reconnaissant la fonction génératrice de Z évaluée en $t = 1$.

c) Soit $X_1, X_2, \ldots$ une suite de variables i.i.d. positives à moments finis. Expliquer pourquoi on peut dire que, pour une grande valeur de n, la variable

$$\sqrt[n]{X_1 \cdots X_n}$$

suit approximativement une loi log-normale dont vous préciserez les paramètres.

$$Z_n := \ln\sqrt[n]{X_1\cdots X_n} = \frac{\ln X_1 + \cdots + \ln X_n}{n}$$

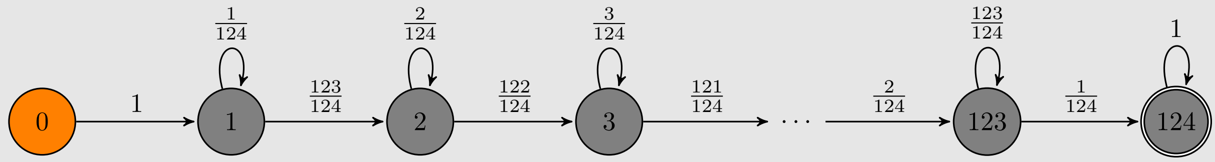
est la moyenne des variables i.i.d. $\ln X_i$, disons d'espérance commune μ et variance σ^2. Par le théorème central limite, on sait que lorsque n grand, Z_n est approximativement normale de paramètres μ et $\frac{\sigma^2}{n}$. On en conclut que X_n est approximativement log-normale de paramètres μ et $\sigma/\sqrt{n}$.

Remarque : on n'a pas en général de relation simple entre l'espérance et la variance des X_i et des $\ln X_i$.

Exercice 3

a) Les supermarchés *Aufrais* proposent ce printemps des autocollants *Super Gontrand* selon la quantité d'achats effectués. Sachant que l'album complet compte 124 autocollants que l'on suppose équiprobables, combien d'autocollants doit-on se procurer en moyenne pour compléter l'album ? (modéliser la situation à l'aide d'une chaîne de Markov à 125 états)

On considère une chaîne de Markov dont l'état i correspond au fait de disposer dans la collection de i autocollants distincts et chaque transition à l'obtention d'un nouvel autocollant.



En effet : une fois qu'on a i autocollants distincts dans la collection, la probabilité d'obtenir l'un des $124-i$ autocollants souhaités au prochain tirage de dépend pas du nombre total d'autocollants déjà obtenus (invariance temporelle).

Si l'on note Y_i le nombre de transitions nécessaires pour compléter la collection en étant à l'état i, on a

$$Y_i = 1 + \frac{i}{124}\,Y_i + \left(1 - \frac{i}{124}\right)Y_{i+1}, \qquad 0 \leqslant i \leqslant 123, \qquad Y_{124}(n) = 0.$$

En prenant l'espérance de part et d'autre, on trouve

$$\mathbb{E}[Y_i] = \frac{1}{1 - \frac{i}{124}} + \mathbb{E}[Y_{i+1}], \qquad 0 \leqslant i \leqslant 123$$

(ou de façon équivalente : on peut considérer le passage de l'état i vers l'état $i+1$ comme une loi géométrique avec probabilité de succès $1 - \frac{i}{124}$), d'où

$$\mathbb{E}[Y_0] = 124\left(\frac{1}{124} + \frac{1}{123} + \cdots + \frac{1}{3} + \frac{1}{2} + 1\right) \approx 670.$$

b) En réalité les autocollants viennent en pochettes de 3 : adaptez votre modélisation de la question a) pour en tenir compte en prenant soin de bien expliciter les hypothèses effectuées.

Plusieurs réponses sont possibles en fonction des hypothèses effectuées... Par exemple :

(a) si on suppose que chaque pochette est constituée de 3 autocollants pris au hasard *avec remise*, c'est équivalent à effectuer 3 transitions (indépendantes) successives dans la chaîne précédente, et ne fait donc que diviser le nombre de transitions nécessaires par 3 (en prenant la partie entière supérieure à la fin de la chaîne) ;

(b) si on suppose que chaque pochette est constituée de 3 aucollants pris au hasard *sans remise*, ça change les probabilités de transition :

$$\mathbb{P}[X_i \to X_{i+k}] = \frac{\binom{124-i}{k}\binom{i}{3-k}}{\binom{124}{3}} \qquad \text{pour } k = 0, 1, 2, 3$$

(c) si on suppose qu'il n'y a qu'un petit nombre $N < \binom{124}{3}$ de pochettes différentes disponibles, on peut tâcher de mettre en place une modélisation plus fine (et donc compliquée) de la situation.

c) Parmi les 124 autocollants, on en compte 12 brillants et 12 en tissu qui sont plus rares. En supposant que seulement 10 % des autocollants distribués sont rares (et que ceux-ci sont équiprobables entre eux), reprendre encore une fois la modélisation en a).

Le plus simple est de considérer que les états correspondent aux valeurs des 3 compteurs : $i \in [\![0, 100]\!]$ de cartes normales, $j \in [\![0, 12]\!]$ de cartes brillantes et $k \in [\![0, 12]\!]$ de cartes en tissu. On obtient donc une chaîne de Markov à $100 \cdot 12^2 = 14400$ états pour lesquels

$$\mathbb{P}[(i,j,k) \to (i+1,j,k)] = \frac{9}{10} \cdot \frac{100-i}{100},$$

$$\mathbb{P}[(i,j,k) \to (i,j+1,k)] = \frac{1}{10} \cdot \frac{12-j}{12}, \qquad \mathbb{P}[(i,j,k) \to (i,j,k+1)] = \frac{1}{10} \cdot \frac{12-k}{12}$$

et $\mathbb{P}[(i,j,k) \to (i,j,k)]$ le complémentaire.

On ne demande pas de calculer à la main la durée moyenne de parcours pour passer de $(0,0,0)$ à $(100,12,12)$! Mais on pourrait le faire en MATLAB sans trop de mal. Intuitivement, la réponse devrait être supérieure à celle de a) puisque les cartes rares sont plus difficiles à obtenir.

Exercice 4

La *loi de Pareto* se rencontre dans plusieurs contextes (par exemple la distribution des salaires ou de la richesse) très asymétriques dans lesquels une petite proportion d'observations peut prendre des valeurs très élevées. Plus précisément, une variable aléatoire X suit une loi de Pareto de paramètres m et $\alpha > 0$ lorsque

$$\mathbb{P}[X > x] = \left(\frac{m}{x}\right)^\alpha \qquad \text{pour tout } x \geqslant m.$$

a) Calculer et représenter graphiquement la fonction de répartition $F(x)$ et la densité $f(x)$ de X.

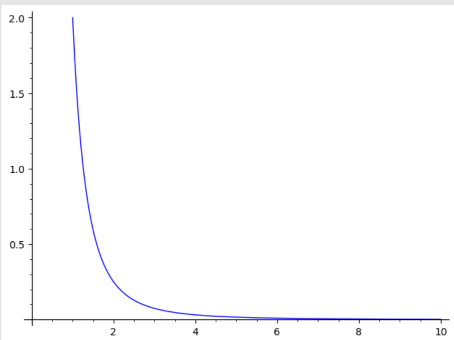
On remarque : $\mathbb{P}[X > m] = 1$ donc la distribution est supportée sur $[m, +\infty[$. Pour $x \geqslant m$,

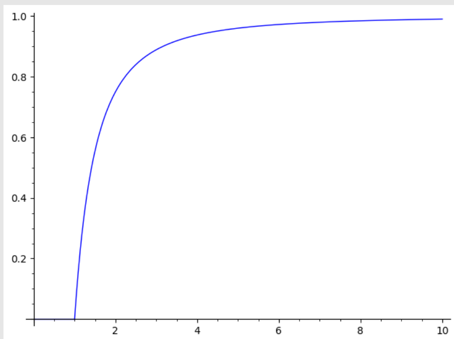
$$F_X(x) = \mathbb{P}[X \leqslant x] = 1 - \left(\frac{m}{x}\right)^\alpha$$

et donc

$$f_X(x) = F_X'(x) = \frac{\alpha\, m^\alpha}{x^{\alpha+1}}.$$

Ci-dessous l'allure de f_X et F_X pour $m = 1$, $\alpha = 2$.





b) En supposant $\alpha > 1$, montrer que $\mathbb{E}[X] = \dfrac{\alpha}{\alpha-1} m$.

Que se passe-t-il lorsque $0 < \alpha \leqslant 1$?

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f_X(x)\,\mathrm{d}x = \int_m^\infty \frac{\alpha\, m^\alpha}{x^\alpha}\,\mathrm{d}x = -\frac{\alpha\, m^\alpha}{\alpha-1}\left[\frac{1}{x^{\alpha-1}}\right]_m^\infty = \frac{\alpha\, m^\alpha}{\alpha-1}\left(0 - \frac{1}{m^{\alpha-1}}\right) = \frac{\alpha}{\alpha-1} m.$$

Lorsque $\alpha \leqslant 1$: l'intégrale impropre ci-dessus diverge, l'espérance de X est infinie.

c) Pour estimer le paramètre m de la loi, on pourrait utiliser la variable aléatoire

$$M = \min(X_1, \ldots, X_n)$$

calculée à partir d'un n-échantillon $X_1, \ldots, X_n$ i.i.d. Pareto (m, α). Vérifier que M suit une loi de Pareto de paramètres m et $n\alpha$ et en déduire un estimateur sans biais de m (en supposant α connu).

$$\mathbb{P}[M > x] = \mathbb{P}[X_1 > x \text{ et } \cdots \text{ et } X_n > x] = \mathbb{P}[X_1 > x] \cdots \mathbb{P}[X_n > x] = \left(\frac{m}{x}\right)^{n\alpha}$$

par indépendance, et on reconnaît la description d'une v.a. suivant une loi Pareto de paramètres m et $n\alpha$. On pourrait utiliser M comme estimateur de m mais comme

$$\mathbb{E}[M] = \frac{n\alpha}{n\alpha - 1} m,$$

cet estimateur est biaisé et aura tendance à surestimer m. On corrige ce biais en utilisant plutôt

$$\widetilde{M} := \frac{n\alpha - 1}{n\alpha} M$$

pour lequel on a $\mathbb{E}[\widetilde{M}] = m$.

Statistiques

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x)\, \mathrm{d}x, \qquad \mathrm{Var}(X) = \sigma_X^2 = \mathrm{Cov}(X, X)$$

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y], \qquad \mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X\, \sigma_Y}$$

Fonction génératrice

$$g_X(t) = \widehat{f_X}\left(-\frac{t}{2\pi j}\right) = \mathbb{E}[e^{\,tX}] = \int_{-\infty}^{+\infty} f_X(x)\, e^{tx}\, \mathrm{d}x = \sum_{n=0}^{\infty} \frac{\mu_n}{n!} t^n = 1 + \mu_1\, t + \frac{\mu_2}{2} t^2 + \frac{\mu_3}{6} t^3 + \cdots$$

avec $\mu_n = \mathbb{E}[X^n]$ les moments de la variable aléatoire X et f_X sa densité

Quelques lois fréquemment utilisées

- loi binomiale $\mathcal{B}(n, p)$:

$$\mathbb{P}[X = k] \;=\; \binom{n}{k} p^k (1-p)^{n-k} \quad (k = 0, 1, \ldots, n)$$

espérance $\;np$, $\qquad$ variance $\;np(1-p)$

fonction génératrice $\;(1 - p + pe^t)^n$

- loi géométrique $\mathcal{G}(p)$:

$$\mathbb{P}[X = k] \;=\; (1-p)^{k-1} p \quad (k \geqslant 1)$$

espérance $\;\dfrac{1}{p}$, $\qquad$ variance $\;\dfrac{1-p}{p^2}$

$$\text{fonction génératrice} \quad \frac{p}{e^{-t}+p-1}$$

- loi de Poisson $\mathcal{P}(\lambda)$:

$$\mathbb{P}[X=k] = \frac{\lambda^k}{k!}e^{-\lambda} \quad k \geqslant 0$$

$$\text{espérance} \quad \lambda, \qquad \text{variance} \quad \lambda$$

$$\text{fonction génératrice} \quad e^{\lambda(e^t-1)}$$

- loi uniforme $\mathcal{U}([0,1])$:

$$\text{densité} \quad f(x) = 1 \quad (0 \leqslant x \leqslant 1)$$

$$\text{espérance} \quad \frac{1}{2}, \qquad \text{variance} \quad \frac{1}{12}$$

$$\text{fonction génératrice} \quad \frac{e^t-1}{t}$$

- loi exponentielle $\mathcal{E}(\lambda)$:

$$\text{densité} \qquad f(x) = \lambda e^{-\lambda x} \quad (x \geqslant 0)$$

$$\text{espérance} \quad \frac{1}{\lambda}, \qquad \text{variance} \quad \frac{1}{\lambda^2}$$

$$\text{fonction génératrice} \quad \frac{\lambda}{\lambda-t}$$

- loi normale $\mathcal{N}(\mu,\sigma^2)$:

$$\text{densité} \quad f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\text{espérance} \quad \mu, \qquad \text{variance} \quad \sigma^2$$

$$\text{fonction génératrice} \quad e^{\mu t}\, e^{\frac{1}{2}\sigma^2 t^2}$$

Fonction de répartition d'une loi normale centrée réduite