Cursus Intégré
2004-2005

Rappels de statistique mathématique
Enoncé des travaux dirigés n°1 à 8

Guillaume Lacôte
Bureau **E03**
✉ Guillaume.Lacote@ensae.fr
☞ http://ensae.no-ip.com/SE222/

Version du 20050421-15h38, révisée le 21 avril 2005

Table des matières

1 Travaux Dirigés n°1

Enoncé de l'exercice 1

On dispose d'observations Y_i relatives au comportement de remboursement ou de non-remboursement d'emprunteurs :

$$Y_i = \begin{cases} 1 & \text{si l'emprunteur } i \text{ rembourse son crédit} \\ 0 & \text{si l'emprunteur } i \text{ est défaillant} \end{cases}$$

Afin de modéliser ce phénomène, on suppose l'existence d'une variable aléatoire Y_i^* normale, d'espérance m et de variance σ^2, qu'on appellera "capacité de remboursement de l'individu i", telle que :

$$Y_i = \begin{cases} 1 & \text{si } Y_i^* \geq 0 \\ 0 & \text{si } Y_i^* < 0 \end{cases}$$

☞ Q1 On note Φ la fonction de répartition de la loi $\mathcal{N}(0,1)$.
Exprimer la loi de Y_i en fonction de Φ.

☞ Q2 Les paramètres m et σ^2 sont-ils identifiables ?

Enoncé de l'exercice 2

Un système S 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.

☞ Q1 Montrer que

$$X \rightsquigarrow \mathcal{E}(\lambda) \Leftrightarrow \forall x \in \mathbb{R},\ \mathbb{P}(X > x) = \exp(-\lambda x)$$

☞ Q2 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.

☞ Q3 On dispose de n systèmes $S_1, ..., S_n$ identiques dont on observe les durées de vie $Z_1, ...Z_n$.

(a) Ecrire le modèle statistique correspondant et la vraisemblance des observations.
A-t-on suffisamment d'information pour estimer λ_1 et λ_2 ?

(b) Si on observe à la fois les durées de vie des systèmes et la cause de la défaillance (machine 1 ou 2), écrire le modèle statistique correspondant et la vraisemblance des observations.
A-t-on alors suffisamment d'information pour estimer λ_1 et λ_2 ?

☞ Q4 Dans cette question, on considère un seul système S utilisant une machine de type 1 et u machine de type 2, mais on suppose que l'on dispose d'un stock de n_1 machines de type 1, durées de vie $X_1^1, ..., X_1^{n_1}$, et d'un stock de n_2 machines de type 2, de durées de vie $X_2^1, ..., X$ Quand une machine tombe en panne, on la remplace par une machine du même type, tant q le stock de machines de ce type n'est pas épuisé. Quand cela arrive, on dit que le système lui-même est en panne. On note toujours Z la durée de vie du système.
Le cas $n_1 = n_2 = 0$ correspond donc à la première question (pas de stock).

(a) Donner la loi de la somme de n variables indépendantes qui suivent une loi exponentie de même paramètre λ.

(b) Ecrire Z en fonction des X_j^i et en déduire $P(Z \geq t)$ en fonction de certaines lois gamm dont on précisera les paramètres.

On note alors N le nombre de machines (des deux types) sorties du stocks quand système tombe en panne, et Z_0 la durée écoulée avant la première panne d'une machi On note Z_i la durée écoulée entre la i-ème panne et la $(i+1)$-ème panne. La durée de totale du système est donc :

$$Z = \sum_{i=0}^{N} Z_i$$

La $(N+1)$-ème panne est donc la panne fatale au système.

(c) Montrer que les variables Z_i sont i.i.d. et donner leur loi.
On pourra utiliser (après l'avoir démontré) le résultat suivant :
Si X est une variable aléatoire de loi exponentielle de paramètre λ, alors

$$\mathbb{P}(X \geq s+t | X \geq s) = \mathbb{P}(X \geq t) = e^{-\lambda t}$$

(on dit que X est "sans mémoire").

(d) Préciser l'ensemble des valeurs possibles pour la variable N et en donner la loi.

(e) On admet que N et les Z_i sont indépendantes. Calculer $\mathbb{E}(Z|N)$ en fonction de N, λ_1 λ_2.
Donner l'expression de $\mathbb{E}(Z)$ en fonction de $\mathbb{E}(N)$, λ_1 et λ_2.

Enoncé de l'exercice 3

Ecrire la vraisemblance et déterminer une statistique exhaustive pour un échantillon de observations i.i.d. de lois :

☞ Q1 loi de Poisson de paramètre λ :

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!},\ \ k \in \mathbb{N}^*$$

☞ Q2 loi de Pareto de paramètres α et θ avec $\alpha > 1$, $\theta > 0$ de densité :

$$f(x) = \frac{\alpha - 1}{\theta} \left(\frac{\theta}{x}\right)^{\alpha} \mathbb{1}_{[\theta;+\infty[}(x)$$

☞ Q3 loi de Weibull de paramètre α et θ avec $\alpha > 0, \theta > 0$ de densité :

$$f(x) = \alpha\theta x^{\alpha-1} e^{-\theta x^{\alpha}} \mathbb{1}_{[0;+\infty[}(x)$$

☞ Q4 loi uniforme sur $[0, \theta]$ avec $\theta > 0$ inconnu.

Enoncé de l'exercice 4

On veut compter le nombre θ de poissons dans un lac fermé. Pour cela, on tire un poisson au hasard, on le marque et on le remet dans le lac. On tire un second poisson. S'il est déjà marqué, on en prend note et on le remet dans le lac. Sinon, on le marque à son tour et on le remet dans le lac. Et ainsi de suite.

On tire n poissons selon la procédure ci-dessus. Au n-ième tirage, l'observation consiste en une variable aléatoire Y_n qui vaut 1 si le poisson est déjà marqué, 0 sinon. Par définition, on a $Y_1 = 0$. Le but de l'exercice est de montrer que :

$$R_n = \sum_{i=1}^{n} Y_i$$

est une statistique exhaustive pour θ.

☞ Q1 Montrer que :

$$\mathbb{P}(Y_n = y_n, ..., Y_1 = y_1) = \begin{cases} & \mathbb{P}(Y_n = y_n | Y_{n-1} = y_{n-1}, \ldots, Y_1 = y_1) \\ \times & \mathbb{P}(Y_{n-1} = y_{n-1} | Y_{n-2} = y_{n-2}, ..., Y_1 = y_1) \\ \vdots & \\ \times & \mathbb{P}(Y_1 = y_1) \end{cases}$$

☞ Q2 Montrer que la loi conditionnelle de Y_n sachant $R_{n-1} = r_{n-1}$ est une Bernoulli de paramètre :

$$\frac{n - r_{n-1} - 1}{\theta}$$

et en déduire que la vraisemblance est proportionnelle à :

$$\prod_{i=1}^{n} \frac{(\theta - i + 1 + r_{i-1})^{1-y_i}}{\theta} \tag{1}$$

☞ Q3 Montrer que l'expression (1) se réécrit :

$$\frac{1}{\theta^n} \frac{(\theta - 1)!}{(\theta - n - 1 + r_n)!}$$

☞ Q4 En déduire que R_n est une statistique exhaustive pour θ.

---

2 Travaux Dirigés n°2

Enoncé de l'exercice 1

On considère n systèmes dont les durées de vie $X_1, \ldots, X_n$ suivent indépendamment une même loi de densité f. On observe uniquement les durées de vie $Y_1, ..., Y_r$ des r premiers systèmes tombant en panne.

☞ Q1 Ecrire la loi du r-uplet $(Y_1, ..., Y_r)$, puis celle de la variable Y_r.

☞ Q2 On suppose ici que la loi des X_i est exponentielle de densité

$$f(x) = \frac{1}{\theta} \exp\left(\frac{-(x-\alpha)}{\theta}\right) \mathbb{1}_{[\alpha,+\infty[}(x)$$

où $\theta > 0$ et $\alpha \geq 0$ sont des paramètres inconnus.
Ecrire la loi du r-uplet $(Y_1, ..., Y_r)$ dans ce cas.

☞ Q3 Trouver une statistique exhaustive pour les paramètres α et θ.

Enoncé de l'exercice 2

Calculer l'information de Fisher dans les modèles statistiques suivants :

☞ Q1 une loi de Poisson de paramètre λ :

$$\mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \quad \text{pour } k \in \mathbb{N}$$

☞ Q2 une loi de Pareto de paramètres α et θ avec $\alpha > 1$ et $\theta > 0$, de densité :

$$f(x) = \frac{\alpha - 1}{\theta} \left(\frac{\theta}{x}\right)^{\alpha} \mathbb{1}_{x \geq \theta}$$

☞ Q3 une loi de Weibull de paramètres α et θ avec $\alpha > 0$ et $\theta > 0$ de densité :

$$f(x) = \alpha\theta x^{\alpha-1} e^{-\theta x^{\alpha}}$$

☞ Q4 loi uniforme sur $[0, \theta]$ avec $\theta > 0$ inconnu.

Enoncé de l'exercice 3

On dispose de n observations $y_1, \ldots, y_n$ sur les durées de vie de certains composants
dustriels. On suppose que les variables aléatoires $Y_1, \ldots, Y_n$ associées sont i.i.d, de dens
$f(t) = \frac{1}{\theta} e^{-\frac{t}{\theta}} \mathbb{1}_{t \geq 0}$.

☞ Q1 Soit F la fonction de répartition des Y_i. On cherche à estimer la "fonction de survie" de cha
composant $\bar{F}(t) = 1 - F(t)$.
Calculer $\bar{F}(t)$ en fonction de t et θ.

☞ Q2 Calculer l'estimateur du maximum de vraisemblance de θ et en déduire un estimateur converg
$\hat{F}(t)$ de $\bar{F}(t)$. Que peut-on dire du biais de $\hat{F}(t)$?

☞ Q3 Calculer la loi limite de $\sqrt{n}\left(\hat{F}(t) - \bar{F}(t)\right)$.

Soit T la variable aléatoire définie par :

$$T = \begin{cases} 1, \text{ si } Y_1 > t \\ 0, \text{ sinon.} \end{cases}$$

On note par ailleurs $S(Y) = Y_1 + \cdots + Y_n$.

☞ Q4 (a) Déterminer la loi de Y_1 conditionnellement à S.

(b) Calculer

$$T^*(Y) = \mathbb{E}(T|S(Y))$$

Comment s'appelle cet estimateur ?

(c) Montrer que T^* est l'estimateur sans biais de $\bar{F}(t)$ optimal (parmi les estimateurs s
biais).

(d) Peux-on dire si T^* est efficace (à distance finie) ?

Enoncé de l'exercice 4

On étudie une variable aléatoire X, de densité $f(\cdot, \theta)$ (f est C^1).

☞ Q1 Quelle est la fonction score du modèle, notée $S(X;\theta)$?
Donner l'expression de l'information de Fisher $I_X(\theta)$.

☞ Q2 En fait, on ne parvient pas à observer X, mais seulement Y définie par :

$$Y = \begin{cases} 1, \text{ si } X \geq s \\ 0, \text{ si } X < s \end{cases}$$

où s est un seuil connu.

On suppose que l'on peut intervertir $\int_{\mathcal{X}}$ et $\frac{\partial}{\partial\theta}$; donner la fonction score du modèle, no

$S_Y(y;\theta)$.
En déduire que

$$S_Y(y;\theta) = \mathbb{E}\left(S_X(X;\theta)|Y=y\right)$$

☞ Q3 En déduire alors que $I_X(\theta) >> I_Y(\theta)$, où $I_Y(\theta)$ est l'information de Fisher associée à Y (l'inégalité s'entend au sens des matrices symétriques).
Quelle interprétation pouvez-vous donner à l'inégalité ci-dessus ?

---

3 Travaux Dirigés n°3

Enoncé de l'exercice 1

Soit Y un vecteur aléatoire de taille N et X une matrice aléatoire à N lignes et K colonne
Soit $\theta \to S(\theta)$ une application de classe C^1 définie sur un voisinage de 0 dans $\mathbb{R}$, à valeurs da
l'ensemble des matrices symétriques définies positives de taille N. On suppose que $S(0) =$
et on note : $A = \frac{\partial S}{\partial \theta}(0)$.
On rappelle par ailleurs que

$$\begin{aligned} \frac{\partial \ln(|S|)}{\partial \theta} &= Tr\left(S^{-1}S'\right) \\ \frac{\partial S^{-1}}{\partial \theta} &= -S^{-1}S'S^{-1} \end{aligned}$$

On considère le modèle linéaire (conditionnel) gaussien suivant :

$$\mathbb{P}_Y^X = \mathcal{N}(Xb, S(\theta))$$

où $b \in \mathbb{R}^K$ et $\theta \in \mathbb{R}$ sont les paramètres inconnus.

☞ Q1 Ecrire la vraisemblance du modèle.

☞ Q2 Ecrire le vecteur score (de taille $K+1$) et vérifier qu'il est d'espérance nulle.

☞ Q3 Calculer la matrice d'information du modèle en $(b, \theta = 0)$.
Le résultat s'exprime très simplement en fonction de X et A.

☞ Q4 Dans cette question, on suppose que $S(\theta)$ est la matrice de terme général :

$$S(\theta)_{i,j} = \left(\theta^{|i-j|}\right)_{1\le i,j\le N}$$

Calculer $I_{b,\theta}(b,0)$.

Enoncé de l'exercice 2

☞ Q1 Calculer l'estimateur du maximum de vraisemblance (e.m.v.) $\hat{p}$ de p dans le modèle

$$X_i \underset{iid}{\rightsquigarrow} \mathcal{B}(1,p)$$

et calculer la loi limite de $\sqrt{n}(\hat{p}-p)$.

☞ Q2 Calculer l'e.m.v. $(\hat{m}, \hat{\sigma}^2)$ de (m, σ^2) dans le modèle

$$X_i \underset{iid}{\rightsquigarrow} \mathcal{N}\left(m, \sigma^2\right)$$

et donner la loi limite du vecteur

$$\sqrt{n}\begin{pmatrix} \hat{m} - m \\ \hat{\sigma}^2 - \sigma^2 \end{pmatrix}$$

☞ Q3 Calculer l'e.m.v. $(\hat{a}, \hat{b})$ de (a, b) dans le modèle

$$X_i \underset{iid}{\rightsquigarrow} \mathcal{U}([a,b]) \quad \text{loi uniforme sur } [a,b]$$

Donner la loi limite du vecteur

$$n\begin{pmatrix} \hat{a} - a \\ b - \hat{b} \end{pmatrix}$$

Enoncé de l'exercice 3

Exemple tiré de *Basu D.(1988) Stastical Information and Likelihood, Springer-Verlag, N.Y.*
Dans une urne contenant 1000 tickets, 20 sont marqués θ et 980 sont marqués 10θ.

☞ Q1 Donner l'estimateur du maximum de vraisemblance $\widehat{\theta}$ de θ lorsque l'on tire un unique ticket de valeur X, et montrer que $P(\widehat{\theta} = \theta) = 0.98$.

☞ Q2 On renumérote les tickets marqués 10θ par $a_i\theta$ $(1 \leq i \leq 980)$ où les a_i sont des réels connus, deux-à-deux distincts, et compris dans l'intervalle $[10, 10.1]$. Donner le nouvel estimateur du maximum de vraisemblance $\widetilde{\theta}$ et montrer que $P(\widetilde{\theta} = \theta) = 0.02$. Ce résultat vous semble-t-il paradoxal ?

---

4 Travaux Dirigés n°4

Enoncé de l'exercice 1

Soit f la densité de la loi exponentielle de paramètre $\frac{1}{\theta}$ translatée de α

$$f(x) = \frac{1}{\theta}\exp\left[-\frac{x-\alpha}{\theta}\right]\mathbb{1}_{[\alpha,+\infty[}(x)$$

☞ Q1 Donner les e.m.v. $\hat{\alpha}$ et $\hat{\theta}$ de α et θ.

☞ Q2 Calculer la loi (à distance finie) de $n(\hat{\alpha} - \alpha)$.

☞ Q3 Déterminer la loi limite de $\sqrt{n}(\hat{\theta} - \theta)$.

☞ Q4 Rappeler l'expression de la loi de la statistique d'ordre $(X_{(1)}, \ldots, X_{(N)})$ en fonction de f. déduire la loi du n-uplet

$$(nX_{(1)}, (n-1)(X_{(2)} - X_{(1)}), \ldots, 2(X_{(n-1)} - X_{(n-2)}), X_{(n)} - X_{(n-1)})$$

En déduire que $\hat{\theta}$ et $\hat{\alpha}$ sont indépendants (à distance finie).

Enoncé de l'exercice 2

Cet exercice présente les bases de l'estimation bayésienne.

☞ Q1 Soit $(\Omega, \mathcal{A}, \mathcal{P})$ un espace probabilisé, A un événement non-vide et $(H_1, \ldots, H_n)$ un *systè complet d'hypothèses incompatibles non-vides* (c'est-à-dire une partition de Ω).
Exprimer $\mathbb{P}(H_i|A)$ en fonction des probabilités de $H_1, \ldots, H_n$, de celles de A conditionnellem à H_i et inversement, mais pas de $\mathbb{P}(A)$.

☞ Q2 Soit $(X_1, \ldots, X_n) \underset{iid}{\rightsquigarrow} \mathcal{L}_\theta$, de paramètre inconnu $\theta \in \Theta$.
On suppose que θ suit une loi *a priori* de densité π_0 sur Θ.
Donner la densité $\pi_{\cdot|x}$ de la loi *a posteriori* de θ conditionnellement à l'observation de $X = x$.

☞ Q3 Définissons pour tout estimateur $\widehat{\theta}(x)$ de θ la *fonction de risque quadratique*

$$R_\nu\left(\widehat{\theta}\right) = \int_\Theta \mathbb{E}_\theta\left(\widehat{\theta}(X) - \theta\right)^2 d\nu(\theta))$$

où ν désigne une loi quelconque sur Θ (par exemple $d\nu = \pi_0 d\lambda$).
On appelle *estimateur bayésien* de θ l'estimateur $\widehat{\theta}$ qui minimise le risque associé.
Montrer que l'estimateur bayésien de θ associé au risque R_{π_0} est

$$\widehat{\theta}(x) = \mathbb{E}_{\pi_{\cdot|x}}(\theta) = \int_\Theta \theta\pi_{\cdot|x}(\theta)d\theta$$

☞ Q4 Donner l'estimateur bayésien de $\theta \in [a,b]$ lorsque $\pi_0 = \mathcal{U}_{[a,b]}$ en fonction de la densité de la loi $\mathcal{L}_\theta$ de X.
Expliciter cet estimateur lorsque $\mathcal{L}_\theta$ est la loi exponentielle $\mathcal{E}(\theta)$ de paramètre inconnu $\theta > 0$.

Enoncé de l'exercice 3

Un des premiers exemples d'utilisation de la Statistique Bayésienne remonte à Laplace, en 1786. Celui-ci décida de répondre à la question suivante : au regard du nombre observé n_g de naissances masculines parmi n naissances à Paris, peut-on dire si la probabilité p qu'un enfant qui naisse soit un garçon est supérieure à $\frac{1}{2}$?

☞ Q1 Laplace munit le paramètre p d'une loi a priori uniforme sur l'intervalle $[0,1]$. Ce choix vous semble-t-il naturel ?

☞ Q2 Exprimer alors la loi a posteriori de p, puis exprimer la probabilité $\mathbb{P}\left(p > \frac{1}{2} | N_g = n_g\right)$ sous forme d'un rapport d'intégrales.
Remarque : Laplace, obtint, pour $n = 493472$ et $n_g = 251527$, une probabilité $\mathbb{P}\left(p > \frac{1}{2} | N_g = n_g\right) \simeq 1 - 1.15.10^{-42}$ et en conclut que p était très vraisemblablement plus grand que $\frac{1}{2}$.

☞ Q3 Donner l'espérance et la variance de la loi a posteriori en fonction de la vrai valeur p_0 du paramètre p. Quelles sont leurs limites lorsque $n \to +\infty$?
Rappel : la loi Beta $\mathcal{B}(\alpha,\beta)$ de paramètres $\alpha > 0$, $\beta > 0$ admet pour densité :

$$f_{\alpha,\beta}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \mathbb{1}_{[0,1]}(x)$$

où $B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$.
Son espérance est $\mathbb{E}(X) = \frac{\alpha}{\alpha+\beta}$ et sa variance $\mathbb{V}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

☞ Q4 Ces limites sont-elles modifiées si on choisit pour loi a priori sur p une loi $\mathcal{B}(\alpha,\beta)$ quelconque ?
Pourquoi est-il judicieux malgré tout de se restreindre au moins à la classe des lois a priori telles que $\alpha = \beta$?

---

5 Travaux Dirigés n°5

Enoncé de l'exercice 1

On souhaite évaluer et analyser le phénomène du chômage. Pour cela, on dispose de n obs vations sur les durées $y_i, 1 \leq i \leq n$, pendant lesquelles des individus sont restés sans emploi
On suppose dans la suite que les variables aléatoires correspondantes $(Y_i)_{i \in [\![1,n]\!]}$ sont i.i.d. suivent la loi de Weibull de paramètres a et b. On rappelle que cette loi est continue sur $\mathbb{R}^+$ admet la fonction de répartition pour $y > 0$

$$F(y;a,b) = 1 - \exp\left(-ay^b\right)$$

On définit la fonction de survie par

$$S(y) = 1 - F(y)$$

et la fonction de hasard par $h(y) = \frac{f(y)}{S(y)}$.

■ Partie 1 **Généralités**

☞ Q1 Donner l'expression de la fonction de hasard du modèle.

☞ Q2 Quelle est en terme de chômage l'interprétation de la fonction de hasard ?
Expliquer alors pourquoi il est important de considérer le cas particulier où cette fonction constante.
Pour quelles valeurs des paramètres, la fonction de hasard est-elle constante ?
Quelles sont alors les lois des durées de chômage ?

☞ Q3 Etudier l'évolution de la fonction de hasard en fonction de a, puis en fonction de b.

■ Partie 2 **Estimation contrainte**

On suppose dans cette partie $b = 1$. Le modèle est alors uniquement paramétré par a.

☞ Q1 Le modèle est-il exponentiel ? Si oui, expliciter une statistique exhaustive.

☞ Q2 Déterminer le vecteur du score et vérifier directement qu'il est centré.

☞ Q3 Quel est l'estimateur du maximum de vraisemblance $\hat{a}_0$ de a ?
Est-il sans biais, y a-t-il surestimation ou sous-estimation systématique ?

☞ Q4 Déterminer la variance asymptotique de cet estimateur $\hat{a_0}$

■ Partie 3 **Estimation non contrainte**

On considère maintenant le cas où a et b peuvent a priori prendre toutes valeurs positives.

☞ Q1 Le modèle est-il exponentiel avec une statistique exhaustive dont la taille est indépendante nombre n d'observations ?
Si oui, expliciter une telle statistique.

☞ Q2 Ecrire les équations de vraisemblance. Sont-elles résolubles sous forme analytique ?

☞ Q3 Donner la forme de la variance asymptotique de l'estimateur du maximum vraisemblance $\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix}$ du paramètre $\begin{pmatrix} a \\ b \end{pmatrix}$.

☞ Q4 En déduire la variance asymptotique de $\hat{a}$ lorsque $b = 1$.
Comparer alors les estimateurs $\hat{a}$ et $\hat{a}_0$ lorsque $b = 1$.
Quelle conclusion en tirer ?

☞ Q5 Quelle démarche pourrait-on proposer pour étudier la distribution de l'estimateur $\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix}$ lorsque l'échantillon est de petite taille, par exemple $n = 10$?

■ Partie 4 **Cas indépendant - non équidistribué**

On considère maintenant le cas de T observations $Y_1, \ldots, Y_T$ indépendantes, de lois respectives :

$$F(y; e^{\alpha t}, 1) \ , \ t \in [\![1, T]\!] \ , \ \alpha \in \mathbb{R}$$

☞ Q1 Déterminer la vraisemblance du modèle $\mathcal{L}$, et vérifier qu'elle est concave en α à $(y_1, \ldots, y_T)$ fixé.
En déduire l'équation caractérisant l'estimateur du maximum de vraisemblance $\widehat{\alpha_T}$ de α.

☞ Q2 On note $u_t = y_t - e^{-\widehat{\alpha_T} t}$.
Donner l'interprétation de u_t.

☞ Q3 Montrer que l'équation de la vraisemblance correspond à la condition d'orthogonalité de $(u_1, \ldots, u_T)$ et de $1, \ldots, T$ pour un certain produit scalaire que l'on précisera.

Enoncé de l'exercice 2

On étudie entre les dates 0 et T un groupe de n individus sans emploi à la date 0, et on cherche à modéliser les durées de chômage $(T_i)_{i \in [\![1,n]\!]}$.
En pratique, on observe les durées de chômage en mois. Plus précisément, on ne dispose pas de la variable continue T_i, mais seulement de la variable discrète T_i^* donnée par

$$T_i^* = \lceil T_i \rceil$$

Autrement dit, la variable T_i^* vaut $t+1$ si l'individu i a retrouvé du travail entre le t-ème et $(t+1)$-ième mois.
En outre entre t et $t+1$, on suppose que :
– l'individu i reçoit N_t^i offres d'emploi, où N_t^i est une suite de variables i.i.d. de loi de Poisson $\mathcal{P}(\lambda)$;

– si l'individu i est toujours au chômage à la date t, et si parmi les N_t^i offres qu'il reçoit, l'u au moins offre un salaire supérieur à une constante ξ_i, propre à l'individu (appelée salaire réserve), alors l'individu n'est plus au chômage : $(T_i^* = t+1)$;
– les salaires des offres d'emploi sont tirés indépendamment des dates d'arrivée des offres et leur nombre dans une loi de fonction de répartition F.

On suppose dans un premier temps pour simplifier que $T = +\infty$.

☞ Q1 Calculer $\mathbb{P}(T_i^* = t+1 \,|T_i > t)$ en fonction de F, ξ_i, λ .

☞ Q2 En déduire la vraisemblance de $(T_1^*, \ldots, T_n^*)$.

☞ Q3 On suppose que tous les individus ont le même salaire de réserve : $\xi_i = \xi$ et que ce salaire réserve commun ξ est connu, ainsi que la fonction de répartition F.
Trouver l'estimateur du maximum de vraisemblance de λ.
Montrer directement que cet estimateur est convergent et asymptotiquement efficace qua $n \to +\infty$.

☞ Q4 En pratique, l'enquête se termine à la fin du T-ième mois, $T < +\infty$: à cette date, certa individus sont encore au chômage ; on n'observe donc que :

$$\begin{aligned} T_i^{**} &= T_i^* \text{ si } T_i^* \leq T \\ &= T+1 \text{ si } T_i^* > T \end{aligned}$$

Ecrire la vraisemblance des observations $(T_1^{**}, \ldots, T_n^{**})$.
Donner l'estimateur du maximum de vraisemblance de λ.
Remarque : On suppose toujours le salaire de réserve commun ξ et la fonction de répartition connue.

☞ Q5 On suppose désormais que F s'écrit :

$$F(x) = \left(1 - e^{-\gamma(\xi - \xi_0)}\right) \mathbb{1}_{\xi \geq \xi_0}$$

où γ est un paramètre inconnu à estimer et ξ_0 est connu.
Le couple (λ, γ) est-il identifiable ?

Enoncé de l'exercice 3

■ Partie 1 **Préliminaire**

☞ Q1 On rappelle que la loi exponentielle de paramètre λ admet la densité $f(y, \lambda) = \begin{cases} \lambda e^{(-\lambda y)}, \text{ si } y \\ 0 \text{ sinon.} \end{cases}$
Soient n variables aléatoires $Y_1 ... Y_n$ i.i.d. de densité $f(\cdot, \lambda)$.
Montrer que $\sum_{i=1}^{n} Y_i$ suit une loi de densité $\lambda^n \frac{y^{n-1}}{(n-1)!} e^{-\lambda y} \mathbb{1}_{[0,+\infty]}(y)$ (on pourra commencer le montrer pour $n = 2$ puis procéder par récurrence).

Dans tout le problème, Z_i et C_i, pour $i \in [\![1,n]\!]$ désignent des variables aléatoires indépendantes suivant des lois exponentielles de paramètres respectifs $\lambda, \mu \in \mathbb{R}^{+*}$.

■ Partie 2 **Observation parfaite**

On dispose d'un échantillon d'observations $(z_i, c_i)_{i\in[\![1,n]\!]}$

☞ Q1 Ecrire le modèle statistique correspondant.
S'agit-il d'une famille exponentielle ? Si oui, peut-on exhiber une statistique exhaustive ?

☞ Q2 Quel est l'estimateur du maximum de vraisemblance $\begin{pmatrix} \hat{\lambda} \\ \hat{\mu} \end{pmatrix}$ du paramètre $\begin{pmatrix} \lambda \\ \mu \end{pmatrix}$?

☞ Q3 Déterminer la loi asymptotiquement de $\begin{pmatrix} \hat{\lambda} \\ \hat{\mu} \end{pmatrix}$.

☞ Q4 Déterminer la loi (à distance finie) de $\begin{pmatrix} \hat{\lambda} \\ \hat{\mu} \end{pmatrix}$.
Est-il biaisé à distance finie ?
Calculer sa matrice de variance-covariance.

☞ Q5 Proposer des estimateurs sans biais optimaux de λ et μ, si possible efficaces.

■ Partie 3 **Observation imparfaite**

On suppose dans cette seconde partie que les seules observations disponibles portent sur $X_i = \text{Min}(Z_i, C_i)$, $i = 1, ..., n$.

☞ Q1 Calculer la fonction de répartition de la variable X_i pour $i \in [\![1,n]\!]$

☞ Q2 Ecrire le modèle statistique correspondant et déterminer les fonctions identifiables du paramètre $\begin{pmatrix} \lambda \\ \mu \end{pmatrix}$.

☞ Q3 Quels sont les estimateurs du maximum de vraisemblance de $\gamma = \lambda + \mu$ fondés :
i) sur $(X_i)_{i\in[\![1,n]\!]}$;
ii) sur $(Z_i, C_i)_{i\in[\![1,n]\!]}$?
Est-il naturel que ces estimateurs soient différents ?

☞ Q4 Comparer les propriétés asymptotiques de ces estimateurs.

■ Partie 4 **Conclusion**

☞ Q1 Déduire des parties **I** et **II** l'expression de l'espérance conditionnelle

$$\text{E}\left(\frac{1}{\sum_{i=1}^n \min(Z_i, C_i)} \middle| \sum_{i=1}^n Z_i, \sum_{i=1}^n C_i \right)$$

6 Travaux Dirigés n°6

Enoncé de l'exercice 1

On considère un échantillon $X_1, \ldots, X_n$ i.i.d. tiré de la loi de Poisson de paramètre λ.

☞ Q1 Donner $\mathbb{E}(X_i)$ et $\mathbb{V}(X_i)$.
Calculer $\phi(t) = \mathbb{E}\left(e^{tX_i}\right)$.
En déduire que

$$\begin{aligned} \mathbb{E}(X) &= \lambda \\ \mathbb{E}\left(X^2\right) &= \lambda(1+\lambda) \\ \mathbb{E}\left(X^3\right) &= \lambda\left(1+3\lambda+\lambda^2\right) \\ \mathbb{E}\left(X^4\right) &= \lambda\left(1+7\lambda+6\lambda^2+\lambda^3\right) \end{aligned}$$

☞ Q2 (a) On pose $\widehat{\lambda_1}(x) = \frac{1}{n}\sum_{i=1}^n x_i = \bar{x}$ et $\widehat{\lambda_2}(x) = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$.
Par quelle méthode d'estimation ont été obtenus $\widehat{\lambda_1}$ et $\widehat{\lambda_2}$?

(b) Les estimateurs $\widehat{\lambda_1}$ et $\widehat{\lambda_2}$ estiment-ils λ sans biais ?
Proposer un autre estimateur sans biais de λ, que l'on notera $\widehat{\lambda_3}$.

(c) Donner la loi (à distance finie) de $\widehat{\lambda_1}$.

(d) Donner la loi asymptotique jointe de $\begin{pmatrix} \widehat{\lambda_1} \\ \widehat{\lambda_2} \end{pmatrix}$, puis celle de $\begin{pmatrix} \widehat{\lambda_1} \\ \widehat{\lambda_3} \end{pmatrix}$.

(e) Soit $\widehat{\lambda_\infty} = \alpha\widehat{\lambda_1} + \beta\widehat{\lambda_3}$, $(\alpha, \beta) \in \mathbb{R}^{+2}$.
Donner la valeur de (α, β) telle que $\widehat{\lambda_\infty}$ soit le meilleur estimateur asymptotiquement sa
biais de λ.
Ce résultat est-il surprenant (donner l'e.m.v. de λ) ?

☞ Q3 Dans cette question, on se place dans le cas où $n = 2$.

(a) Calculer les erreurs quadratiques moyennes pour les trois estimateurs $\widehat{\lambda_1}$, $\widehat{\lambda_2}$ et $\widehat{\lambda_3}$.

(b) Déterminer en fonction de λ lequel de ces estimateurs est le meilleur selon la critère
l'erreur quadratique moyenne.

(c) Comparer explicitement l'EQM de $\hat{\lambda}_3$ avec la borne FDCR.

Enoncé de l'exercice 2

On considère une population de n individus infectés par un virus ; on étudie leurs dur
d'incubation $(T_i)_{i\in[\![1,n]\!]}$, dont on suppose qu'elle est observable.

Pour modéliser l'hétérogénéité de la population, on suppose qu'on peut caractériser chaque individu i par un "facteur de risque" inobservable, réalisation de la variable aléatoire Λ_i, de telle sorte que :

- la loi de T_i, conditionnellement à Λ_i est la loi exponentielle de paramètre λ_i (de densité $\lambda_i e^{-\lambda_i t}\mathbb{1}_{t\geq 0}$) ;
- la famille $(\Lambda_i)_{i\in[\![1,n]\!]}$ est identiquement distribuée selon la loi $\mathcal{L}$ de densité

$$f(\lambda) = \frac{\alpha^r}{\Gamma(r)}\lambda^{r-1}e^{-\alpha\lambda}\mathbb{1}_{\mathbb{R}^+}$$

  où $\alpha > 0$ et $r > 2$;
- les couples (T_i, Λ_i) sont indépendants entre eux.

☞ Q1 Donner la vraisemblance de $(T_1, \ldots, T_n)$.

☞ Q2 Calculer, lorsqu'il existe, le moment d'odre k $\mathbb{E}\left(T_i^k\right)$.

☞ Q3 Calculer l'information de Fisher du modèle.
Dans le cas où α est connu, calculer l'estimateur du maximm de vraisemblance de r.
Que se passe-t-il si α et r sont tous deux inconnus ?

☞ Q4 On suppose α connu.
Déterminer au moyen de la méthode des moments un estimateur convergent de r.
Cet estimateur est-il sans biais ? Est-il asymptotiquement efficace ?

☞ Q5 On suppose α et r inconnus.

(a) En utilisant les deux premiers moments de T_i, trouver des estimateurs convergents $\tilde{\alpha}$ et $\tilde{r}$ de α et r.

(b) Donner la loi limite du vecteur

$$\sqrt{n}\begin{pmatrix} \bar{T} - \mathbb{E}(T) \\ \bar{T^2} - \mathbb{E}(T)^2 \end{pmatrix}$$

En déduire la loi asymptotique du vecteur $(\tilde{\alpha}, \tilde{r})$.

Enoncé de l'exercice 3

On considère les réalisations de T variables aléatoires i.i.d. $Y_1, Y_2, \ldots, Y_T$, issues d'une loi de Poisson de paramètre λ inconnu.

On s'intéresse au test de l'hypothèse :

$$H_0 : \text{“ } \lambda = \lambda_0\text{” contre } H_a : \text{“}\lambda \neq \lambda_0\text{”}$$

où λ_0 est un réel donné.

☞ Q1 Pour tester ce type d'hypothèse, on dispose de trois tests asymptotiques usuels : test de **Wa** du **score** et du **rapport de maximum de vraisemblance**. Rappeler rapidement leur princ et définir les statistiques de test sur lesquelles ils s'appuient.

☞ Q2 Pratiquer explicitement chacun de ces tests, en calculant l'expression de la statistique de t ξ_T^{test} et en définissant la région critique au seuil $\alpha \in [0,1] : W_\alpha = \{(y_1, \ldots, y_T) \,/\, \xi_T^{test}(y_1, \ldots, y_T$ $q_{1-\alpha}^{\chi_1^2}\}$

☞ Q3 On veut tester $\lambda_0 = 1$ au seuil $\alpha = 5\%$.
Décider de chacun des tests dans le cas où l'échantillon observé $(y_1, \ldots, y_T)$ est tel que :
- $T = 100$ et $\bar{y} = 1,2$;
- $T = 200$ et $\bar{y} = 1,1$.

Discuter de l'acceptation de H_0 et expliquer pourquoi certains tests ne conduisent pas à même décision.
Conclure.

---

7 Travaux Dirigés n°7

Enoncé de l'exercice 1

Pour étudier l'arrivée des appels dans un central téléphonique, on comptabilise lors de 200 observations consécutives, le nombre d'appels observés par seconde, ce qui produit les résultats suivants :

Nombre d'appels par seconde	Effectifs observés
0	6
1	15
2	40
3	42
4	37
5	30
6	10
7	9
8	5
9	3
10	2
11	1

On suppose les arrivées des appels indépendantes, et en outre que la probabilité élémentaire λdt qu'il arrive un appel entre les instants t et $t+dt$ est indépendante de la date t.

Autrement dit

- le nombre N_t d'appels observés sur l'intervalle de temps $[0,t]$ suit une loi de Poisson de paramètre λt.
- plus généralement $N_{t+s} - N_t$ est indépendant de N_t et suit une loi de Poisson de paramètre λs.

☞ Q1 (a) Tester l'adéquation de la loi empirique à la famille des lois de Poisson.

(b) Quel est le niveau limite permettant d'accepter l'adéquation à la loi de Poisson ?

☞ Q2 A une date antérieure le nombre d'appels sur un intervalle de temps $[0,t]$ suivait une loi de Poisson de paramètre 4.
Tester la stabilité du comportement entre les deux dates.

☞ Q3 Que donnerait un test d'adéquation à une loi de Poisson de paramètre $3,7$? Conclure.

Table de la loi de Poisson :

	$\mathbb{P}_{3,7}(X=k)$	$\mathbb{P}_4(X=k)$
0	0.0247	0,0183
1	0.0915	0,0733
2	0.1684	0,1465
3	0.2087	0,1957
4	0.1930	0,1954
5	0.1428	0,1563
6	0.0881	0,1042
7	0.0465	0,0595
8	0.0215	0,0298
9	0.0099	0,0132
10	0.0046	0,0053
11	0.0021	0,0019
12	0.0001	0,0006
13	< 0.0001	0,0002
14	< 0.0001	0,0001

Enoncé de l'exercice 2

On s'intéresse à la proportion des ménages équipés d'un magnétoscope. Pour cela, on t
de manière équiprobable avec remise un échantillon de n ménages, et on observe pour chac
ménage $i \in [\![1,n]\!]$ la variable

$$y_i = \begin{cases} 1 \text{ si le ménage } i \text{ est équipé} \\ 0 \text{ sinon.} \end{cases}$$

☞ Q1 Réaliser un test de l'hypothèse nulle : "la proportion des ménages équipés n'excède pas 20 %".

☞ Q2 On se demande si la probabilité qu'un ménage i soit équipé n'est pas fonction d'une varia
(scalaire) x_i donnée (revenu, âge du chef de famille...).
On définit à cet effet un modèle statistique conditionnel à X_i de la façon suivante :

$$\mathbb{P}(y_i = 1 \mid x_i) = \frac{e^{a+bx_i}}{1+e^{a+bx_i}}$$

où a et b sont deux paramètres réels inconnus.
On cherche alors à tester $\underline{H_0 :}\ b = 0$.

(a) Calculer le score et la matrice d'information de Fisher du modèle pour n observations.

(b) Expliciter le modèle contraint par H_0.
Calculer les estimateurs du maximum de vraisemblance $\widehat{a^0}$ et $\widehat{b^0}$ de a et b dans ce modèle, puis évaluer le score en $\left(\widehat{a^0}, \widehat{b^0}\right)$.

(c) Donner un estimateur $\widehat{I_1^{H_0}}$ de I_1, convergent sous H_0 et fonction des estimateurs contraints des paramètres.

(d) Exprimer la statistique du test du score de l'hypothèse H_0.
Quelle est sa loi asymptotique sous H_0 ? Sur quelle corrélation repose le test ?

Enoncé de l'exercice 3

Un industriel reçoit K lots de n pièces, avec la garantie que la proportion p de pièces défectueuses est la même dans chaque lot et inférieure à 5%. Pour vérifier que la garantie est exacte, on tire avec remise des pièces dans chaque lot, jusqu'à obtenir une pièce défectueuse par lot.
Soit Y_k, la variable aléatoire désignant le nombre de tirages nécessaires dans le lot k.

☞ Q1 Calculer la loi de Y_k.
Calculer $\mathbb{E}(Y_k)$ et $\mathbb{V}(Y_k)$.

☞ Q2 Proposer un test de Wald de l'hypothèse :

$$\underline{H_0 : p = 5\%}$$

☞ Q3 Soit $A_k = \left\{\bar{Y} < k_\alpha\right\}$, pour k_α donné ; calculer $\mathbb{P}_p(A_k)$ en fonction de $p = p_0$.
Montrer que $\sup_{p \leq 5\%} \left(\lim_{K\infty} \mathbb{P}_p(A_k)\right) = \lim_{K\infty} \mathbb{P}_{p=5\%}(A_k)$.
En déduire un test asymptotique de l'hypothèse :

$$\underline{H_0 : p \leq 5\%}$$

Enoncé de l'exercice 4

On dispose de n observations i.i.d. d'un couple de variables aléatoires positives scalaires (W_i, X_i).
Le but de l'exercice est de suggérer une procédure pour tester l'hypothèse H_0 selon laquelle la loi conditionnelle de W_i sachant X_i est une loi de Pareto, de densité :

$$f(W|X,b) = \frac{bX}{W_0}\left(\frac{W_0}{W}\right)^{bX+1} \mathbb{1}_{W \geq W_0}$$

pour $b > 0$ et $W_0 > 0$

☞ Q1 On suppose dans un premier temps que les X_i sont tous égaux à $X \in \mathbb{R}$ connu, et que W_0 connu, de sorte que le seul paramètre inconnu du modèle est $\alpha = bX + 1$.

(a) Calculer l'espérance et la variance de W, notées m et v.
A quelle condition sur α, supposée vérifiée par la suite, les moments m et v existent-ils ?

(b) Déterminer l'estimateur du maximum de vraisemblance $\widehat{\alpha}$ de α.

(c) Déterminer le score et l'information de Fisher du modèle.
En déduire $\mathbb{E}(\ln W)$ et $\mathbb{E}((\ln W)^2)$.

(d) Soit $s = \frac{1}{n}\sum_{i=1}^n \left(\ln(W_i) - \overline{\ln W}\right)^2$, où $\overline{\ln W}$ est la moyenne empirique des logarithmes W_i.
On se propose de fonder un test de H_0 sur la différence :

$$s - \frac{1}{\left(\widehat{\alpha} - 1\right)^2}$$

Justifier un tel test.

(e) Ecrire le Théorème Central Limite pour $\left(\ln W_i, (\ln W_i)^2\right)_{i \in [\![1,N]\!]}$, et tester l'hypothèse
(On donnera la forme de la matrice de variance-covariance sans pour autant la calcu explicitement).

☞ Q2 On suppose toujours W_0 connu, mais les X_i prennent désormais des valeurs a priori distinct

(a) Calculer l'estimateur du maximum de vraisemblance de b.

(b) Montrer que : $\mathbb{E}\left(X^2\left(\ln\frac{W_0}{W}\right)^2 - \frac{2X}{b}\ln\frac{W_0}{W}\right) = 0$.
Comment testeriez-vous alors l'hypothèse H_0 ?

☞ Q3 On suppose enfin que W_0 est inconnu.

(a) Calculer l'estimateur du maximum de vraisemblance de W_0.

(b) Peut-on adapter la procédure de test précédemment mise-en-œuvre ?

---

8 Travaux Dirigés n°8

Enoncé de l'exercice 1

Soit un échantillon $X_1, \dots, X_n$ de variables aléatoires i.i.d telles que :
– X_i a une probabilité α de valoir a, et
– une probabilité $(1-\alpha)$ de suivre une loi normale $\mathcal{N}(0,1)$,
ce qui s'écrit encore :

$$X_1 = a \cdot \mathbb{1}_{Z_1=1} + Y_1 \cdot \mathbb{1}_{Z_1=0}$$

où $Z_1 \rightsquigarrow \mathcal{B}(1,\alpha)$ est indépendante de $Y_1 \rightsquigarrow \mathcal{N}(0,1)$.

☞ Q1 Montrer que $(\delta_a + \lambda)^{\otimes n}$ est une mesure dominante pour le modèle considéré.

☞ Q2 Montrer que :

$$\frac{dP_{X_1,\dots,X_n}(x_1,\dots,x_n)}{d(\delta_a+\lambda)^{\otimes n}} = \prod_{i=1}^{n} (\alpha \mathbb{1}_a(x_i) + (1-\alpha)(1-\mathbb{1}_a(x_i))\phi(x_i))$$

où ϕ est la densité d'une loi normale $\mathcal{N}(0,1)$.

Enoncé de l'exercice 2

Cet exercice s'inspire librement des travaux de Dov Samet, 2003.[1]

On s'intéresse pour tout couple (X,Y) de variables aléatoires réelles à la propriété $\mathcal{P}$ suivante :

$$(\mathcal{P}) \quad \forall \mathcal{A} \text{ mesurable}, \begin{cases} \mathbb{P}(X < Y \text{ et } X \in \mathcal{A}) = \mathbb{P}(X > Y \text{ et } X \in \mathcal{A}) = \frac{1}{2}\mathbb{P}(X \in \mathcal{A}) \\ \mathbb{P}(Y < X \text{ et } Y \in \mathcal{A}) = \mathbb{P}(Y > X \text{ et } Y \in \mathcal{A}) = \frac{1}{2}\mathbb{P}(Y \in \mathcal{A}) \end{cases}$$

☞ Q1 Comment s'interprête la propriété $\mathcal{P}$?

■ Partie 1 ***"Devinez quel est le plus grand !"***

On s'intéresse au problème suivant, présenté originellement par Blackwell (1951) :

> Deux nombres réels sont tirés aléatoirement, et chacun est placé dans une enveloppe. L'une d'elle est (indépendamment) tirée au hasard et vous est présentée. Vous devez deviner, au vu du nombre qu'elle contient, s'il s'agit du plus grand ou le plus petit des deux. Sauriez-vous deviner juste plus d'une fois sur deux en moyenne ?

[1] Résultats présentés à la 14ième conférence internationale de Théorie des jeux, Stony Brook 2003.
Voir `Http ://www.sunysb.edu/gametheory/Conf03/twopuzzles.pdf` pour un exposé plus complet.

Pour simplifier la présentation, supposons que vous gagnez +1 si vous avez deviné juste et sinon ; ce problème revient à savoir si vous pouvez vous garantir un gain espéré strictem positif.

☞ Q1 Soient X et Y les deux nombres tirés aléatoirement, et supposons que (X,Y) vérifie $\mathcal{P}$. Quel est votre meilleur gain espéré ? Interprêter.

☞ Q2 Cherchons néanmoins à construire une stratégie gagnante.
Pour ce faire, donnons-nous $s \in \mathbb{R}$ et soit $\sigma_s : \begin{pmatrix} \mathbb{R} & \to & \{-1,1\} \\ x & \mapsto & \begin{cases} 1 & \text{si } x > s \\ -1 & \text{sinon} \end{cases} \end{pmatrix}$ la *stratégie seuil* associée à s.

(a) Soit $(x \neq y)$ une réalisation du tirage ; le nombre qui vous est présenté est soit x, soit Quel est votre gain espéré en jouant selon σ_s sachant que $x < s$ et $y < s$?

(b) Quel est votre gain espéré sachant que $x < s < y$?

(c) En déduire que si votre gain espéré (sur tous les tirages possibles) est strictement positif.

☞ Q3 En déduire qu'aucune paire de variables aléatoires réelles (X,Y) ne peut vérifier $\mathcal{P}$.

■ Partie 2 ***Le jeu des enveloppes***

On s'intéresse désormais au célèbre jeu dit *des enveloppes*, introduit par Kraitchik (1953) s la forme suivante :

> On considère deux enveloppes contenant une certaine somme d'argent, mais l'une contenant le double de l'autre ; chacune a autant de chances que l'autre de contenir la plus grosse somme. L'une d'elles vous est donnée au hasard, et vous devez choisir entre la prendre, ou prendre l'autre.
> A supposer qu'elle contienne x, l'autre a une chance sur deux de contenir $2x$ et une chance sur deux de contenir $\frac{x}{2}$; donc le contenu espéré de l'autre enveloppe est $\frac{5}{4}x > x$, de sorte que vous avez intérêt à changer d'enveloppe. Mais le même raisonnement s'applique aussi à la nouvelle enveloppe, de sorte que vous avez à nouveau intérêt à changer d'enveloppe, et ainsi de suite ...
> Que faire dans cette situation ?

Résoudre ce paradoxe a été le prétexte à une littérature foisonnante ; voici une solution.
Notons X la somme (variable aléatoire réelle) contenue dans l'enveloppe qui vous est présent et Y celle contenue dans l'autre.

☞ Q1 Quel sens donner à l'expression $\mathcal{H}_x$: "l'autre a une chance sur deux de contenir $2x$ et une cha sur deux de contenir $\frac{x}{2}$" ?

☞ Q2 En déduire que si cette hypothèse est vraie, alors la distribution de (X,Y) vérifie $\mathcal{P}$.

☞ Q3 Conclure.

Enoncé de l'exercice 3

Soient $X_1, \ldots, X_n$ n variables aléatoires i.i.d dans $\mathbb{R}^k$, de loi paramétrique $\mathcal{P}_\theta$ de densité

$$\frac{d\mathcal{P}_\theta(x)}{d\lambda} = f(x-\theta)$$

$\theta \in \mathbb{R}^k$ est appelé *paramètre de position*.
On cherche à étudier les estimateurs T du paramètre θ. On se restreint pour cela à la classe des estimateurs *équivariants*, c'est-à-dire des estimateurs T vérifiant :

$$\forall c \in \mathbb{R}^k, T(X_1+c, \ldots, X_n+c) = T(X_1, \ldots, X_n) + c$$

Cette restriction est naturelle : en effet, pour tout $c \in \mathbb{R}^k$, $X_i + c \rightsquigarrow \mathcal{P}_{\theta+c}$, et on ne peut admettre qu'un simple changement d'échelle puisse mener à une estimation différente.

☞ Q1 Montrer que :

$$\text{T équivariant} \Leftrightarrow \begin{array}{l} \exists T_1 : (\mathbb{R}^{k-1} \to \mathbb{R}^k) \ / \ \forall (x_1, \ldots, x_n) \in \mathbb{R}^k \\ T(x_1, \ldots, x_n) = T_1(x_2 - x_1, \ldots, x_n - x_1) + x_1 \end{array}$$

☞ Q2 Soit $W(x,y) = w(x-y)$ une fonction de perte. Montrer que, si un estimateur T est équivariant, le risque associé, pour la perte W, ne dépend pas de θ.
Un estimateur équivariant T qui minimise le risque $R_w(T,0)$ (c'est-à-dire au point $\theta = 0$) est alors un estimateur optimal (parmi les équivariants) pour la fonction de perte W.

☞ Q3 Soit la fonction

$$\psi(x) = \int_{\mathbb{R}^k} w(x-u) \prod_{i=1}^n f(X_i - u) du$$

et l'estimateur T^* défini par

$$\psi(T^*) = \inf_{x \in \mathbb{R}^k} \psi(x)$$

T^* est appelé *l'estimateur de Pitman* (on suppose ici que w est telle que l'équation $\psi(x^*) = \inf_{x \in \mathbb{R}^k} \psi(x)$ admette une solution et une seule).

(a) Montrer que T^* est équivariant.

(b) Montrer que, pour toute fonction φ telle que $\mathbb{E}(\|\varphi(X_1, \ldots, X_n)\|) < +\infty$, on a :

$$E_\theta[\varphi(X_1, ..., X_n) | X_2 - X_1, ..., X_n - X_1] = \int_{\mathbb{R}^k} \varphi(X_1 + \theta - u, ..., X_n + \theta - u) \frac{\prod_{i=1}^n f(X_i - u)}{\int_{\mathbb{R}^k} \prod_{i=1}^n f(X_i - v) dv} du$$

(c) Montrer finalement que T^* est optimal pour w.

☞ Q4 Donner l'expression de T^* lorsque $w(x-y) = \|x-y\|^2$.

☞ Q5 Dans ce cas, que vaut T^* lorsque $X_i \rightsquigarrow \mathcal{N}(\theta, 1)$?

Et lorsque $X_i \rightsquigarrow \mathcal{U}_{[-\frac{1}{2}+\theta, \frac{1}{2}+\theta]}$?

Enoncé de l'exercice 4

On considère une loi multinomiale à K modalités ($K \geq 3$) de probabilités $p_1, p_2, \ldots, p_K$. On d
pose de n observations indépendantes issues de cette loi. On notera N_k le nombre d'observati
de la modalité $k \in [\![1, K]\!]$.
On veut tester l'hypothèse : $\underline{H_0 :}$ "$p_1 + p_2 = \frac{1}{2}$"

☞ Q1 Donner les estimateurs du maximum de vraisemblance contraints $(\widehat{p_k^0})_k$ et non contraints $(\hat{p}$
de $p_1, \ldots, p_K$.

☞ Q2 Calculer la statistique ξ^W du test de Wald de l'hypothèse H_0.

☞ Q3 (a) En constatant que $\widehat{p_3^0}$ est asymptotiquement efficace sous H_0, montrer que

$$\mathbb{C}ov_{as}\left(\widehat{p_3^0}, \widehat{p_3} - \widehat{p_3^0}\right) = 0$$

et en déduire que

$$\mathbb{V}_{as}\left(\widehat{p_3} - \widehat{p_3^0}\right) = \mathbb{V}_{as}\left(\widehat{p_3}\right) - \mathbb{V}_{as}\left(\widehat{p_3^0}\right)$$

(b) Calculer la loi limite de $\left(\widehat{p_3^0} - p_3\right)$, et en déduire $\mathbb{V}_{as}\left(\widehat{p_3^0}\right)$.

(c) Donner un estimateur $\mathbb{V}_{as}\widehat{\left(\widehat{p_3} - \widehat{p_3^0}\right)}$ convergent sous H_0 de $\mathbb{V}_{as}\left(\widehat{p_3} - \widehat{p_3^0}\right)$.
En déduire l'expression de la statistique du test d'Hausman de H_0

$$\xi^H = n \frac{\left(\widehat{p_3} - \widehat{p_3^0}\right)^2}{\mathbb{V}_{as}\widehat{\left(\widehat{p_3} - \widehat{p_3^0}\right)}}$$

(d) Vérifier que le test d'Hausman est convergent.

(e) Montrer que les statistiques ξ^W et ξ^H sont asymptotiquement équivalentes sous H_0.

Enoncé de l'exercice 5

La mise en œuvre de nombreuses techniques d'analyse, et notamment la quasi-totalité des te
niques d'estimation et de test de modèles statistiques, nécessitent d'*engendrer* des réalisati
x d'une variable aléatoire réelle X de loi $\mathcal{L}$ donnée.
L'objet de cet exercice est de présenter quelques techniques permettant à un ordinateur, mach
essentiellement déterministe, d'engendrer de telles variables. Il s'agit donc de construire p
toute loi $\mathcal{L}$ une fonction récursive $f^{\mathcal{L}} : \mathbb{N} \to I$ qui énumère des réalisations d'un variable $X \rightsquigarrow$
i.e. telle que la distribution empirique des $\left(f^{\mathcal{L}}(n)\right)_{n \in \mathbb{N}}$ converge vers la vraie distribution F

☞ Q1 Cherchons tout d'abord à engendrer une variable aléatoire **entière** uniforme.
Soit $N \in \mathbb{N}^*$ et considérons $\mathcal{L} = \mathcal{U}_{[\![0,N-1]\!]}$.

(a) Définissons pour $\phi : [\![0, N-1]\!] \to [\![0, N-1]\!]$ et $s \in [\![0, N-1]\!]$

$$f_{\phi,s} : \begin{pmatrix} [\![0,N-1]\!] & \to & [\![0,N-1]\!] \\ t & \mapsto & \phi^t(s) \end{pmatrix}$$

Montrer que $f_{\phi,s}$ est périodique à partir d'un certain rang ; on note $T_{\phi,s}$ sa période.
En quel sens peut-on dire que $f_{\phi,s}$ "génère" une variable uniformément distribuée sur $[\![0, N-1]\!]$?

(b) En quel sens peut-on dire que $f_{\phi,s}$ est d'autant meilleure que $T_{\phi,s}$ est grande ?

(c) Soient $a \in [\![2, N-1]\!]$, $c \in [\![0, N-1]\!]$ et $p \in [\![2, N-1]\!]$ et définissons $\psi_{a,c,p}$: $\begin{pmatrix} \mathbb{N} & \to & [\![0,N-1]\!] \\ n & \mapsto & (an+c) \mod p \end{pmatrix}$.
A quelles conditions $f_{\psi_{a,c,p},s}$ est-elle un bon générateur uniforme sur $[\![0, N-1]\!]$?

☞ Q2 Cherchons à engendrer une variable aléatoire réelle uniforme sur $[0, 1]$.

(a) Soit $(x_n)_{n\in\mathbb{N}} \in [0,1]^{\mathbb{N}}$, et définissons pour $f : [0,1] \to \mathbb{R}$ continue et $n \in \mathbb{N}^*$

$$S_n^x(f) = \frac{1}{n} \sum_{k=1}^{n} f(x_k)$$

On dit de x qu'elle vérifie le *critère de Weyl* si pour toute f continue

$$S_n^x(f) \xrightarrow[n\to+\infty]{} \int_0^1 f$$

Comment s'interprète ce critère ?

(b) Soit $r \in \mathbb{R}$, et définissons $x = (frac(rk))_{k\in\mathbb{N}}$ où $frac(u) = u - E(u) \in [0, 1[$ désigne la partie fractionnaire de $u \in \mathbb{R}$.
Montrer que x vérifie le critère de Weyl *ssi* r est irrationnel.

☞ Q3 On cherche désormais à engendrer des réalisations d'une variable aléatoire de loi quelconque.

(a) Soit $\mathcal{L}$ la loi de densité $f_{\mathcal{L}} = \frac{1}{3}\delta_1 + \frac{1}{4}\delta_2 + \frac{5}{12}\delta_3$.
Tracer la fonction de répartition $F_{\mathcal{L}}$ de $\mathcal{L}$.
En déduire son inverse généralisée

$$F^{-1} : \begin{pmatrix} [0,1] & \to & \mathbb{R} \\ s & \mapsto & \inf\{x \in \mathbb{R} \,/\, F_{\mathcal{L}}(x) \geq s\} \end{pmatrix}$$

Soit $U \rightsquigarrow \mathcal{U}_{[0,1]}$ et définissons $X = F_{\mathcal{L}}^{-1}(U)$; quelle est la loi de X ?

(b) Soit alors $F_{\mathcal{L}}$ la fonction de répartition supposée continue et strictement croissante d'une loi $\mathcal{L}$ quelconque.
On définit de même $X = F_{\mathcal{L}}^{-1}(U)$ pour $U \rightsquigarrow \mathcal{U}_{[0,1]}$; quelle est la loi de X ?

(c) Proposer un algorithme qui génère des réalisations successives d'une variable X de loi exponentielle $\mathcal{E}(\lambda)$ de densité $f_{\mathcal{E}_\lambda}(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,+\infty[}(x)$ pour $\lambda > 0$.

(d) Soit $(X, Y) \rightsquigarrow \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 I_2\right)$.
Soient $R^2 = X^2 + Y^2$ et $\theta = \arctan \frac{Y}{X}$.
Montrer que R^2 et θ sont indépendantes, et donner leurs lois.
En déduire un algorithme de génération d'une variable gaussienne.

(e) Même question si X suit une loi de Weibull $\mathcal{W}(\alpha, \beta)$ de densité $f_{\mathcal{W}(\alpha,\beta)}(x) = \alpha\beta x^{\beta-1} e^{-\alpha x^{\beta}}$

(f) Même question si X suit une loi de Poisson $\mathcal{P}(\lambda)$ de densité $f_{\mathcal{P}(\lambda)}(k) = \frac{\lambda^k}{k!} e^{-\lambda}$.
On pourra comparer $x_n = u_1 \times \cdots \times u_n$, où chaque u_n est une réalisation de $U \rightsquigarrow U_{[0,1}$ $e^{-\lambda}$.