Université des Sciences et Technologies de Lille
U.F.R. de Mathématiques Pures et Appliquées

CAPES-M1 **Année 2010-2011**

Fiche de T.D. n° 5

Ex 1. *Efficacité*

On considère sur le modèle statistique $(\Omega, \mathcal{F}, (P_\theta)_{\theta\in\mathbb{R}})$, un échantillon $X_1, \ldots, X_n$ tel que la loi de X_1 sous P_θ soit $\mathfrak{N}(\theta, 1)$.

1) Vérifiez que $\overline{X}$ est un estimateur efficace de θ.

Ex 2. *Estimateur du maximum de vraisemblance*

Estimer par maximum de vraisemblance le paramètre θ

a) d'une loi de Poisson ;

b) d'une loi géométrique ;

c) d'une loi exponentielle ;

d) de la loi uniforme sur $[0, \theta]$.

Ex 3. *Estimation par maximum de vraisemblance*

Soit $(\Omega, \mathcal{F}, (P_\theta)_{\theta\in\Theta})$ un modèle statistique où $\Theta =]0, +\infty[$. On note $X_1, \ldots, X_n$ un échantillon associé à ce modèle, les X_i étant des v.a. à valeurs dans $]0, 1[$ ayant pour densité sous P_θ la fonction :

$$t \mapsto f(t, \theta) := \theta t^{\theta-1} \mathbf{1}_{]0,1[}(t).$$

On se propose d'estimer θ par maximum de vraisemblance.

1) Explicitez la fonction de vraisemblance $L(x_1, \ldots, x_n, \theta)$ pour des x_i tous dans $]0, 1[$. Montrez qu'elle admet un maximum unique et en déduire l'estimateur T_n de θ par maximum de vraisemblance.

2) Calculez $\mathbf{E}_\theta(\ln X_1)$ après avoir justifié son existence.

3) Déduire de ce qui précède que T_n est un estimateur fortement consistant de θ.

Ex 4. *Estimation de la durée d'une panne*

Des étudiants font un TP en utilisant n ordinateurs reliés à un serveur [1]. Chaque ordinateur envoie au serveur, à des instants aléatoires, des requêtes variées (exécution de commande, transmission de données, etc).

1. Concrètement, un gros ordinateur sans écran installé dans un local verrouillé, climatisé, et sous alarme. Le seul point important ici est que les étudiants ont la possibilité d'utiliser le serveur, mais pas de le voir.

A l'instant $t = 0$ ce serveur tombe en panne. Il redémarre au bout d'une durée τ.

Les étudiants ne savent pas quelle est la durée τ de la panne. Mais pour i de 1 à n, ils constatent au bout de quel temps T_i après le début de panne la première requête de l'ordinateur i est acceptée par le serveur. Pour estimer la durée τ de la panne, ils disposent donc de n données $T_1(\omega), \cdots, T_n(\omega)$.

Les requêtes ayant lieu à des instants aléatoires, il est naturel de supposer que la durée $X_i = T_i - \tau$ entre le redémarrage du serveur et la première requête de l'ordinateur i suit une loi exponentielle. Le paramètre a de cette loi exponentielle sera supposé connu et identique pour toutes les machines (que représente-t-il ?). On supposera également que les v.a. $X_1, \cdots, X_n$ sont indépendantes.

1) Quelle est la loi de la v.a. $V = \inf(X_1, \cdots, X_n)$?

2) Estimer la durée inconnue τ à partir de $T_1, \cdots, T_n$ en utilisant la méthode du maximum de vraisemblance. On note W l'estimateur obtenu.

3) Calculer le biais de W. En déduire un estimateur sans biais Z de la durée de panne. Quelle est la variance de ce nouvel estimateur ?

4) Fabriquer un autre estimateur sans biais de τ, en utilisant cette fois la moyenne empirique $\overline{T}$ des T_i. Quelle est sa variance ?

5) Des deux estimateurs sans biais de τ, lequel est le meilleur ? N'y a-t-il pas contradiction ici avec l'inégalité de Cramér-Rao ? Pourquoi ?

Ex 5.

Soit $(\Omega, \mathcal{F}, (P_\theta)_{\theta \in]0,1[})$ un modèle statistique et $X_1, \ldots, X_n$ un échantillon associé à ce modèle, tel que pour tout $\theta \in]0,1[$, la loi sous P_θ des X_i soit la loi discrète sur $\{a, b, c\}$ donnée par

$$P_\theta(X_i = a) = \frac{1-\theta}{2}, \quad P_\theta(X_i = b) = \frac{1}{2}, \quad P_\theta(X_i = c) = \frac{\theta}{2}.$$

Les valeurs des réels a, b, c n'interviennent pas dans ce problème, la seule chose qui importe est qu'il s'agisse de 3 valeurs distinctes. On définit les statistiques

$$S_{n,a} := \sum_{i=1}^{n} \mathbf{1}_{\{X_i = a\}}, \quad S_{n,b} := \sum_{i=1}^{n} \mathbf{1}_{\{X_i = b\}}, \quad S_{n,c} := \sum_{i=1}^{n} \mathbf{1}_{\{X_i = c\}}.$$

1) Vérifiez que $\frac{2}{n} S_{n,c}$ est un estimateur sans biais et fortement consistant de θ.

2) Que peut-on dire de $S_{n,a} + S_{n,b} + S_{n,c}$? Quelle est la loi du vecteur aléatoire $(S_{n,a}, S_{n,b}, S_{n,c})$?

3) Montrez que

$$\sum_{n=1}^{+\infty} P_\theta(S_{n,b} = n) < +\infty.$$

4) On pose

$$\Omega' := \{\omega \in \Omega;\ \exists N(\omega) \in \mathbb{N},\ \forall n \geq N(\omega),\ S_{n,a}(\omega) + S_{n,c}(\omega) > 0\}.$$

Expliquez pourquoi $P_\theta(\Omega') = 1$ pour tout $\theta \in]0,1[$.

5) On définit la suite de variables aléatoires $(T_n)_{n\geq 1}$ par

$$T_n(\omega) := \begin{cases} \dfrac{S_{n,c}(\omega)}{S_{n,a}(\omega)+S_{n,c}(\omega)} & \text{si } S_{n,a}(\omega)+S_{n,c}(\omega)>0, \\ 0 & \text{sinon.} \end{cases}$$

Montrez soigneusement que T_n est un estimateur fortement consistant de θ.

6) Montrez que T_n est un estimateur asymptotiquement sans biais de θ. Indication : notez que pour tout $\omega \in \Omega$, $T_n(\omega) \in [0,1]$, donc $(T_n)_{n\geq 1}$ est une suite de v.a. bornées par une même constante.

7) Dans l'échantillon observé $x_1 = X_1(\omega), \ldots, x_n = X_n(\omega)$, on note $n_a := S_{n,a}(\omega)$, le nombre d'occurences de la valeur a et de même $n_b := S_{n,b}(\omega)$, $n_c := S_{n,c}(\omega)$. Exprimez à l'aide de ces notations la vraisemblance $L(x_1,\ldots,x_n,\theta)$ de l'échantillon.

8) Montrez que pour $(x_1,\ldots,x_n)$ tel que $n_c>0$, la fonction $\theta \mapsto L(x_1,\ldots,x_n,\theta)$ a un maximum unique atteint en un point $\widehat{\theta} \in]0,1[$ qui s'exprime simplement en fonction de n_a et n_c. En déduire un estimateur de θ par maximum de vraisemblance.

Ex 6. *Un mélange de gaussiennes*

On considère le modèle statistique $\Big(\Omega, \mathcal{F}, (P_\theta)_{\theta\in[0,1]}\Big)$ et une variable aléatoire $X : \Omega \to \mathbb{R}$, dont la loi sous P_θ a pour fonction de répartition F_θ définie par :

$$\forall \theta \in]0,1[,\ \forall x \in \mathbb{R}, \quad F_\theta(x) = (1-\theta)\Phi\Big(\frac{x-\mu}{\sigma}\Big) + \theta\Phi\Big(\frac{x-\mu}{5\sigma}\Big),$$

où Φ est la fonction de répartition de la loi $\mathfrak{N}(0,1)$. Dans tout l'exercice, les paramètres $\mu \in \mathbb{R}$ et $\sigma \in]0,+\infty[$ sont supposés connus.

1) Quelle est la loi de X sous P_0 ? Sous P_1 ? *Indication :* calculez $P_0\Big((X-\mu)/\sigma \leq x\Big)$.

2) Expliquez pourquoi la loi de X sous P_θ admet une densité f_θ et calculez la.

3) Calculez $\mathbf{E}_\theta X$.

4) Calculez $\mathrm{Var}_\theta\, X$.

5) Au vu du calcul précédent, proposez un estimateur T_n fortement consistant de θ, basé sur un échantillon $X_1,\ldots,X_n$ de v.a. indépendantes de même loi que X sous P_θ.

6) Lors d'une séance de T.D. sur machine à laquelle vous avez échappé, l'enseignante a demandé à son groupe de proposer une méthode de simulation de la loi de X sous P_θ. L'étudiant G. Duflair a proposé de définir une variable aléatoire Y par

$$Y := (\sigma Z + \mu)\mathbf{1}_{\{U>\theta\}} + (5\sigma Z + \mu)\mathbf{1}_{\{U\leq\theta\}},$$

où U est une v.a. de loi uniforme sur $[0,1]$ et Z *indépendante* de U est gaussienne $\mathfrak{N}(0,1)$. Les variables U et Z sont fournies facilement par le générateur de nombre aléatoires utilisé. G. Duflair s'apprête à en tirer un programme Scilab, quand l'enseignante lui demande de justifier sa réponse. Pouvez vous l'aider ?

Ex 7. *Intervalle de confiance*

Dans une verrerie industrielle, une chaîne de production fournit des bouteilles vides. On s'intéresse à la masse moyenne m (en grammes) d'une bouteille produite par cette chaîne. Celle-ci peut s'interpréter comme l'espérance d'une variable aléatoire de loi inconnue. Pour estimer m, on prélève un échantillon de 400 bouteilles que l'on pèse une par une. On obtient ainsi les données numériques x_i ($i = 1 \ldots, 400$) où x_i est la masse de la i^{e} bouteille pesée. Pour vous éviter le travail fastidieux d'entrée de ces données dans une calculatrice, on vous fournit les résultats intermédiaires suivants :

$$\sum_{i=1}^{400} x_i = 79882 \text{ g}, \qquad \sum_{i=1}^{400} x_i^2 = 15963824 \text{ g}^2.$$

Proposez un intervalle de confiance au niveau 95% pour m en indiquant clairement les hypothèses faites et les résultats du cours utilisés.

Ex 8. *Podomètre*

Un podomètre est un appareil qui, fixé à la ceinture d'un marcheur, compte ses pas. Il est aussi capable de calculer approximativement la distance parcourue par le marcheur. La distance parcourue à chaque pas par un marcheur est une variable aléatoire d'espérance μ et d'écart type σ (ces quantités seront exprimées en mètres). Ces deux paramètres dépendent du marcheur et du type de terrain. On les supposera constants pour simplifier. Notons X_i la distance parcourue par le marcheur lors du i^{e} pas. Nous supposerons en outre que les X_i sont indépendantes et de même loi que X. Ce que le podomètre peut mesurer exactement est le nombre N de pas effectués. Pour estimer la distance parcourue en N pas, il affiche simplement la valeur du produit $N\mu$. En pratique, le podomètre est réglé en usine avec un certaine valeur par défaut μ_0 en mémoire et chaque utilisateur a la possibilité de la remplacer par sa valeur personnelle μ. La distance parcourue lors des k premiers pas du marcheur est notée :

$$S_k := \sum_{i=1}^{k} X_i \quad (k \in \mathbb{N}^*).$$

1) On suppose dans cette question que $\mu = 0,75$ m et $\sigma = 0,20$ m. Après une randonnée, le podomètre affiche $N = 12764$ pas et estime la distance parcourue à 9573 m. En justifiant votre réponse, proposez un intervalle $[a, b]$ tel que la distance *réellement* parcourue S_N vérifie :

$$P(a \leq S_N \leq b) \simeq 0,99.$$

2) Le manuel de l'utilisateur du podomètre explique comment évaluer μ (marcher sur une distance connue, par exemple un kilomètre entre deux bornes sur une route, et diviser cette distance par le nombre de pas). Par contre, il ne dit rien de σ et on peut seulement lire que la distance calculée par le podomètre est fournie avec une *précision relative* de $\pm 10\%$.

a) Trouvez $\varepsilon(N, \mu, \sigma)$ tel que

$$P\Big(1 - \varepsilon(N, \mu, \sigma) \leq \frac{S_N}{N\mu} \leq 1 + \varepsilon(N, \mu, \sigma)\Big) \simeq 0,99.$$

b) En supposant μ connue et σ inconnu mais tel que $\sigma/\mu \leq 1/2$, critiquez l'affirmation du manuel sur la précision relative.

3) En fait l'utilisateur du podomètre a quelques notions de statistique. Il lui paraît légitime de supposer que les X_i sont gaussiennes de loi $\mathfrak{N}(\mu, \sigma)$ et il applique la procédure suivante pour estimer μ et σ.

- Il marche sur 10 km de borne à borne le long d'une route et note que son podomètre lui indique $n = 13158$ pas. Il en déduit une valeur estimée $\widehat{\mu_n}(\omega)$ de μ. Laquelle ?
- Il enduit ses semelles d'un liquide coloré et effectue 25 pas sur une route goudronnée. Grâce aux traces de ses semelles, il peut mesurer les valeurs observées des v.a. gaussiennes $Z_1, \ldots, Z_{25}$, où Z_i est la distance parcourue lors du i^{e} pas. Il calcule alors la variance empirique de cet échantillon et trouve $0,0085$ m^2. Il détermine ensuite grâce au théorème de Student un intervalle de confiance de la forme $[0, \widehat{\sigma_{25}}(\omega)[$ pour σ, au niveau de confiance 99%. Expliquez comment et donnez la valeur de $\widehat{\sigma_{25}}(\omega)$.

Pourquoi n'utilise-t-il pas l'échantillon $Z_1, \ldots, Z_{25}$ pour estimer μ ?

4) Notre utilisateur voudrait bien déduire de son travail une formule d'encadrement de S_N avec grande probabilité (ici N est quelconque). Il introduit alors les évènements *indépendants* :

$$A := \left\{ N\mu - 2,575\sigma\sqrt{N} \leq S_N \leq N\mu + 2,575\sigma\sqrt{N} \right\} \tag{1}$$

$$B := \left\{ \mu - \frac{2,575\sigma}{\sqrt{n}} \leq \widehat{\mu_n} \leq \mu - \frac{2,575\sigma}{\sqrt{n}} \right\} \quad (n = 13158) \tag{2}$$

$$C := \left\{ \sigma < \widehat{\sigma_{25}} \right\}. \tag{3}$$

a) Justifiez brièvement les *égalités exactes*[2] $P(A) = P(B) = P(C) = 0,99$.

b) En déduire que $P(A \cap B \cap C) \geq 0,97$.

c) En déduire un encadrement de S_N vrai avec une probabilité d'au moins 97%, dont les bornes s'expriment en fonction de N et des valeurs estimées ci-dessus[3] pour μ et σ.

2. En négligeant les erreurs d'approximation numérique dans le calcul de la f.d.r. Φ de la loi $\mathfrak{N}(0,1)$.

3. On ne vous demande pas de justification rigoureuse du fait que l'on peut remplacer les v.a. $\widehat{\mu_n}$ et $\widehat{\sigma_{25}}$ par les valeurs numériques trouvées ci-dessus. En fait, c'est là que l'indépendance de A, B et C serait utile.