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

IS-Math314 **Année 2006–2007**

Fiche n° 5

Ex 1. Test sur les intentions de vote

1) Soit S_n une variable aléatoire sur $(\Omega, \mathcal{F}, (\mathbf{P}_\theta)_{\theta\in[0,1]})$ de loi binomiale $\text{Bin}(n, \theta)$ sous $\mathbf{P}_\theta$.

a) Montrer pour toute constante a, que l'application

$$\theta \mapsto \mathbf{P}_\theta(S_n \geq a)$$

est croissante sur $[0, 1]$ (en θ). *Indication :* on utilisera une méthode de simulation de la loi binomiale.

b) Soit $\theta_0 \in [0, 1]$. En déduire la construction d'un événement de la forme

$$R = \{S_n \geq a\}$$

de telle sorte que $\sup_{\theta\leq\theta_0} \mathbf{P}_\theta(R) \leq 0,05$.

2) Application à la construction d'un test : Avant une élection, un candidat A affirme qu'il est sûr de dépasser le score des 30%. Pour savoir s'il a raison, on interroge un nombre n suffisant de personnes et on note S_n le nombre de personnes parmi les n interogées qui ont l'intention de voter pour le candidat A. Si le nombre S_n dépasse un certain seuil (à déterminer), le candidat A affirmera que son score dépasse 30% en souhaitant que le risque de se tromper n'excède pas 5%.

On introduit alors le paramètre θ représentant la probabilité (inconnue) de voter pour A sur l'ensemble des electeurs et un modèle statistique $(\Omega, \mathcal{F}, (\mathbf{P}_\theta)_{\theta\in]0,1[})$ dans lequel S_n suit une loi binomiale $\text{Bin}(n, \theta)$ sous $\mathbf{P}_\theta$.

On observe une réalisation de S_n et à l'aide de cette observation, on veut décider l'une des deux hypothèses suivantes :

$$H_0 \quad : \quad \theta \leq 0,3 \qquad \text{ou bien si} \qquad H_1 \quad : \quad \theta > 0,3.$$

La règle de décision sera donc du type

– si on observe $S_n(\omega) \geq a$, alors on rejette H_0 (et on accepte H_1)
– sinon, on accepte l'hypothèse H_0,

a étant un seuil à déterminer. Il est déterminé de telle sorte que la probabilité de rejeter H_0 à tort soit contrôlée par un niveau donné (ici 5%). Par exemple, c'est la probabilité pour le candidat A de conclure que $\theta > 0,3$ alors qu'en réalité $\theta \leq 0,3$.

a) Déterminer le seuil a (en fonction de n).

b) On recueille 32% d'intentions de votes auprés de 500 personnes, que pouvez-vous en conclure ?

Ex 2. Les ex-aequo d'un échantillon

1) On se propose de montrer que si X et Y sont deux variables aléatoires réelles *indépendantes* de fonctions de répartition respectives F et G et si F est *continue*, alors $P(X=Y)=0$. Voici quelques indications. On note pour $M\in\mathbb{N}^*$,

$$\Delta_M := \{(x,y)\in[-M,M]^2;\ x=y\} \quad \text{et} \quad \Delta := \{(x,y)\in\mathbb{R}^2;\ x=y\}.$$

Il s'agit de prouver que $P_{(X,Y)}(\Delta)=0$ (pourquoi ?). En remarquant [1] que pour tout $k\in\mathbb{N}^*$,

$$\Delta_M \subset \bigcup_{-Mk\leq j<Mk} \left[\frac{j}{k},\frac{j+1}{k}\right]^2,$$

utiliser la continuité *uniforme* de F sur le *compact* $[-M,M]$ pour montrer que

$$\forall\varepsilon>0,\quad P_{(X,Y)}(\Delta_M)\leq\varepsilon\bigl(G(M)-G(-M)\bigr).$$

En déduire que $P_{(X,Y)}(\Delta_M)=0$, puis que $P_{(X,Y)}(\Delta)=0$.

2) Montrer que si $X_1,\dots,X_n$ est un échantillon de la loi de f.d.r. F continue, presque sûrement il n'y a pas d'*ex-aequo* dans l'échantillon.

3) On suppose que la f.d.r. F est discontinue au point x_0 et on note

$$N_n(x_0) := \operatorname{card}\{i=1,\dots,n;\ X_i=x_0\}.$$

Montrer que $N_n(x_0) \xrightarrow[n\to+\infty]{\text{p.s.}} +\infty$. On donnera un résultat plus précis que cette affirmation.

Ex 3. Calculs sur la f.d.r. empirique

On note $X_1,\dots,X_n$ un échantillon de la loi de f.d.r. F, *i.e.* les X_i sont indépendantes et de même loi de f.d.r. F. On note F_n la fonction de répartition empirique associée à l'échantillon :

$$\forall x\in\mathbb{R},\quad F_n(x)=\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i\leq x\}}.$$

1) Tracer le graphe de F_n pour une réalisation de l'échantillon $X_1,\dots,X_n$ de votre choix (en prenant $n=5$ par exemple).

2) On suppose dans cette question que les X_i sont des v.a. positives (ou plus généralement que $P(X_1<0)=0$). Calculez $\int_0^{+\infty}(1-F_n(t))\,\mathrm{d}t$ et interprétez le résultat.

Ex 4. Polygone de fréquences cumulées croissantes

Soit $X_1,\dots,X_n$ un échantillon, F la f.d.r. de chaque X_i et F_n la fonction de répartition empirique associée à cet échantillon. On note $X_{n:1},\dots,X_{n:n}$ les statistiques d'ordre de l'échantillon, obtenues par le réordonnement croissant de l'échantillon, *i.e.* pour tout $\omega\in\Omega$, $\{X_1(\omega),\dots,X_n(\omega)\}=\{X_{n:1}(\omega),\dots,X_{n:n}(\omega)\}$ et $X_{n:i}(\omega)\leq X_{n:i+1}(\omega)$ pour $i=1,\dots,n-1$. En particulier $X_{n:1}=\min_{1\leq i\leq n}X_i$ et $X_{n:n}=\max_{1\leq i\leq n}X_i$. On posera par commodité $X_{n:0}:=X_{n:1}-1$ (convention non standard). On définit le lissage polygonal G_n de F_n, comme la fonction continue affine par morceaux valant 1 à droite de $X_{n:n}$, valant 0 à gauche de $X_{n:1}-1$, coïncidant avec F_n en chaque $X_{n:i}$ (donc en chaque point de saut de F_n) et interpolant linéairement entre deux sauts consécutifs de F_n.

1. Dessin impératif !

1) Représentez F_n et G_n pour les deux échantillons observés suivants

$$X_1(\omega) = 2;\ X_2(\omega) = 1{,}5;\ X_3(\omega) = 4;\ X_4(\omega) = 0{,}5.$$

$$X_1(\omega') = 3;\ X_2(\omega') = 2;\ X_3(\omega') = 2;\ X_4(\omega') = 1{,}5;\ X_5(\omega') = 4.$$

2) On suppose dans cette question F continue sur $\mathbb{R}$. En utilisant le théorème de Glivenko-Cantelli, montrez que presque-sûrement G_n converge vers F uniformément sur $\mathbb{R}$ quand n tend vers $+\infty$.

3) Donnez un argument général montrant que cette convergence uniforme presque sûre ne peut avoir lieu lorsque F n'est pas continue sur $\mathbb{R}$. Proposez une autre explication de ce résultat négatif en vous basant sur l'exercice 2.

Entraînement supplémentaire facultatif :

Ex 5. Simulation de lois gaussiennes

Soit $M_i = (U_i, V_i)_{i\in\mathbb{N}^*}$ une suite de vecteurs i.i.d de loi uniforme sur $[-1,1]^2$. On définit la variable aléatoire

$$T := \inf\left\{n \in \mathbb{N}^* \ :\ U_n^2 + V_n^2 < 1\right\},$$

avec la convention $\inf\emptyset = +\infty$.

1) Quelle est la loi de T et celle de M_T ?

2) On pose $R^2 = U_T^2 + V_T^2$ et

$$X = U_T\sqrt{\frac{-2\log(R^2)}{R^2}} \quad \text{et} \quad Y = V_T\sqrt{\frac{-2\log(R^2)}{R^2}}.$$

Quelle est la loi du vecteur (X,Y) ? (On pourra utiliser un changement de variable en coordonnées polaires dans le calcul de $E(f(X,Y))$).

Ex 6. Estimation du paramètre de position de la loi log-normale

Extrait de l'examen d'IS (mai 2005) On rappelle qu'une variable aléatoire Y suit la loi gaussienne $\mathfrak{N}(m,\sigma)$, de paramètres $m \in \mathbb{R}$ et $\sigma \in \mathbb{R}_+^*$, si elle admet pour densité la fonction

$$g : \mathbb{R} \to \mathbb{R}, \quad y \longmapsto \frac{1}{\sigma\sqrt{2\pi}}\exp\Big(-\frac{(y-m)^2}{2\sigma^2}\Big).$$

On sait qu'alors Y admet des moments de tout ordre et que

$$\mathbf{E}Y = m, \quad \operatorname{Var} Y = \sigma^2.$$

1) Exprimez $\mathbf{E}(Y^2)$ en fonction des paramètres m et σ.

2) On dit qu'une variable aléatoire X suit la loi log-normale de paramètres m et σ, avec $m \in \mathbb{R}$ et $\sigma \in \mathbb{R}_+^*$, si elle admet pour densité la fonction

$$f : \mathbb{R} \to \mathbb{R}, \quad x \longmapsto \frac{1}{x\sigma\sqrt{2\pi}}\exp\Big(-\frac{(\ln x - m)^2}{2\sigma^2}\Big)\mathbf{1}_{]0,+\infty[}(x).$$

La fonction f est donc nulle sur $\mathbb{R}_-$ et strictement positive sur $\mathbb{R}_+^*$. Vérifiez à l'aide d'un changement de variable que f est bien une densité de probabilité sur $\mathbb{R}$.

3) Vérifiez de même que la variable aléatoire $\ln X$ a des moments de tout ordre et que

$$\mathbf{E}(\ln X) = m, \quad \mathrm{Var}(\ln X) = \sigma^2.$$

4) On se propose dans la suite de l'exercice d'étudier l'estimation du paramètre m. Pour cela on dispose d'un échantillon *observé* $X_1(\omega), \dots, X_n(\omega)$ avec n grand. On considère donc que les variables aléatoires X_i utilisées ici sont indépendantes et de même loi que X et ceci, quelles que soient les valeurs prises par les paramètres m et σ. En utilisant la question précédente, proposez en justifiant votre réponse, un estimateur T_n de m qui soit *fortement consistant*, *i.e.* presque-sûrement convergent vers m lorsque n tend vers $+\infty$.

5) Expliquez brièvement comment construire à partir de la valeur observée $T_n(\omega)$ un intervalle de confiance au niveau $1-\varepsilon$ pour m, en distinguant les deux cas σ connu, puis σ inconnu. Indiquez quels sont les théorèmes limites utilisés pour cela.

6) *(A traiter quand le cours aura avancé)* On suppose que σ est connu et m inconnu. Déterminez l'estimateur de m par la méthode du maximum de vraisemblance.