Université Lille1
Sciences et Technologies

U.F.R. de Mathématiques

MASS 3 S5 Estimation **Année 2013-2014 M.Fradon**

Examen, 16 décembre 2013

Durée : 3 heures.

Matériel autorisé : calculatrice, tables statistiques, feuille manuscrite A4 recto-verso.

Ex 1. On s'intéresse au rendement hebdomadaire d'une action. Ce rendement est défini à partir du cours de cloture S_0 à la fin d'une semaine et du cours de cloture S_1 à la fin de la semaine suivante. Il vaut $\ln(S_1/S_0)$. Par exemple, si l'action cloture à 100€ une semaine et à 101€ la semaine suivante, le rendement est de $\ln(101/100) \simeq +1\%$. Et si elle cloture à 100€ une semaine et à 99€ la semaine suivante, le rendement est de $\ln(99/100) \simeq -1\%$.

Comme il est d'usage avec le modèle de Black-Scholes, on fait l'hypothèse que les rendements successifs sont indépendants et suivent tous la même loi gaussienne. L'espérance de cette loi s'appelle *rendement moyen* et son écart-type porte le nom de *volatilité historique*.

1) On s'intéresse au cours d'une action pendant trois mois. On note X_1, X_2,...,X_{13} ses rendements hebdomadaires pendant ces treize semaines. Construire à partir des X_i deux intervalles de confiance au niveau 90%, l'un pour l'estimation du rendement moyen de l'action et l'autre pour l'estimation de sa volatilité historique.

2) Voici les rendements observés pendant treize semaines pour l'action Total :

$0,0103226723$ $-0,0110303864$ $0,0302172903$ $-0,0181371575$ $0,0167624294$
$0,0266953982$ $-0,0049233623$ $0,0040300066$ $-0,0215693529$ $0,0099955398$
$0,014362904$ $-0,0063705161$ $-0,0382838274$

La moyenne de ces treize nombres vaut vaut $0,0009285875$ et la moyenne de leurs carrés vaut $0,0003678498$. Quels encadrements raisonnables peut-on proposer pour les valeurs du rendement moyen et de la volatilité historique ?

Ex 2. Echantillonage de signal bruité

Une borne Wifi reçoit un signal radio émis par un ordinateur ou un téléphone. Ce signal est en réalité une onde (i.e. une fonction) mais pour simplifier on supposera dans cet exercice que le signal est caractérisé par un nombre réel a.

A cause des obstacles (murs, mobilier et personnes dans la zone de propagation) le signal reçu n'est pas exactement celui qui a été émis : s'y ajoutent des parasites. Lorsque l'ordinateur envoie le signal a, la borne Wifi reçoit un signal $a + W$ où W est une variable aléatoire qui représente les parasites induits par l'environnement. La loi de W n'est pas connue ici, mais il est assez facile de connaitre son espérance et sa variance en mesurant ce que reçoit la borne quand aucun signal n'est émis : $E(W) = 0$ et $\mathrm{Var}(W) = 400$.

1) Un ordinateur émet un signal constant pendant une seconde. La borne enregistre pendant cette seconde 800 valeurs indépendantes $a + W_1$,...,$a + W_{800}$. On veut évaluer la valeur a émise. Construire un intervalle de confiance qui a approximativement 95% de chances de contenir a.

2) Evaluer la précision de l'approximation faite à la question précédente dans le cas où $E(|W|^3) = 1000$. En déduire le niveau de confiance réel de l'intervalle construit.

3) L'ordinateur émet cette fois un signal constant pendant un centième de seconde seulement. La borne n'enregistre pendant cette durée que 8 valeurs indépendantes $a + W_1$, $\ldots$, $a + W_8$. On veut à nouveau évaluer la valeur a émise. Construire un intervalle de confiance au niveau 95% adapté à ce cas. Quel défaut présente cette estimation ?

Ex 3. *Les deux parties de cet exercice sont indépendantes.*

Pour n'importe quel α fixé dans $]0;1[$ on définit la densité de probabilité f_α par

$$\forall x \in \mathbb{R} \quad f_\alpha(x) = \alpha \mathbf{1}_{0\leq x\leq 1} + (1-\alpha)\mathbf{1}_{-1\leq x<0} = (1-\alpha)(\mathbf{1}_{-1\leq x\leq 1}) \left(\frac{\alpha}{1-\alpha}\right)^{(\mathbf{1}_{x\geq 0})}$$

Partie 1

1) Tracer le graphe de f_α. Calculer la fonction de répartition F_α associée à la densité f_α et tracer son graphe.

2) Déterminer la fonction quantile F_α^{-1} et tracer son graphe. Si on connait la valeur numérique de α comment doit-on faire pour réaliser un tirage selon la densité f_α ?

Partie 2

3) Un dispositif nous permet d'effectuer n tirages selon la loi de densité f_α. On construit ainsi les variables aléatoires i.i.d. X_1, X_2,$\ldots$,X_n. On ignore sur quelle valeur numérique de α le dispositif est réglé, et on veut estimer cette valeur. Déterminer l'estimateur $\hat{\alpha}_n$ du maximum de vraisemblance.

4) $\hat{\alpha}_n$ est-il sans biais ? Est-il fortement consistant ? Calculer son erreur quadratique moyenne.

5) Calculer l'espérance et la variance de la variable aléatoire X_1 de densité f_α.

6) Un statisticien propose d'utiliser au lieu de $\hat{\alpha}_n$ un autre estimateur : $\overline{\alpha}_n = \overline{X}_n + \frac{1}{2}$. Prouver que $\overline{\alpha}_n$ est un estimateur sans biais de α et qu'il est fortement consistant. Calculer son erreur quadratique moyenne. Entre $\hat{\alpha}_n$ et $\overline{\alpha}_n$ lequel est le meilleur estimateur ?

Université Lille1 Sciences et Technologies

U.F.R. de Mathématiques

MASS 3 S5 Estimation **Année 2013-2014 M.Fradon**

Corrigé de l'examen du 16 décembre 2013

Ex 1. 1) Les 13 rendements $X_1,\dots,X_{13}$ sont supposés indépendants. Ils suivent la même loi gaussienne, dont la variance est non-nulle car le cours d'une action fluctue réellement de façon non prévisible. On note μ et σ le rendement moyen et la volatilité historique de l'action, qui déterminent la loi gaussienne des X_i. Et on note $\overline{X_{13}}$ et V_{13} leur espérance empirique et variance empirique. D'après le théorème de Student

$$\sqrt{12}\frac{\overline{X_{13}}-\mu}{\sqrt{V_{13}}} \sim \text{Student}(12) \quad \text{et} \quad \frac{13\ V_{13}}{\sigma^2} \sim \chi^2(12)$$

La table de la Student(12) et la symétrie de cette loi donnent :

$$P\left(\left|\sqrt{12}\frac{\overline{X_{13}}-\mu}{\sqrt{V_{13}}}\right| \le 1,782\right) \simeq 90\%$$

On obtient l'intervalle de confiance sur μ : $I = \left[\overline{X_{13}} \pm 1,782\frac{\sqrt{V_{13}}}{\sqrt{12}}\right]$

On utilise ensuite la table du $\chi^2(12)$ qui donne :

$$P\left(5,226 \le \frac{13\ V_{13}}{\sigma^2} \le 21,026\right) = P\left(\frac{13\ V_{13}}{5,226} \ge \sigma^2 \ge \frac{13\ V_{13}}{21,026}\right) \simeq 90\%$$

d'où l'intervalle de confiance sur la volatilité σ : $J = \left[\sqrt{\frac{13\ V_{13}}{21,026}}\ ;\ \sqrt{\frac{13\ V_{13}}{5,226}}\right]$.

2) $\overline{X_{13}}(\omega) = 0,0009285875$ et $V_{13}(\omega) = 0,0003678498 - (\overline{X_{13}}(\omega))^2 \simeq 0,0003669875$ donc

$$I(\omega) \simeq [-0,0089260\ ;\ 0,0107833] \quad \text{et} \quad J(\omega) \simeq [0,015063\ ;\ 0,030215]$$

Ex 2. Echantillonage de signal bruité

1) $V_1 = a + W_1$, …, $V_{800} = a + W_{800}$ sont 800 v.a.i.i.d. d'espérance $E(a+W) = a$ et de variance $\text{Var}(a+W) = 400$ donc d'après le théorème central limite et la table de la loi normale, on a

$$P\left(\left|\sqrt{800}\frac{\overline{V_{800}}-a}{\sqrt{400}}\right| \le 1,96\right) \simeq 95\%$$

c'est-à-dire

$$P\,(a \in I) \simeq 95\% \quad \text{quand} \quad I = \left[\overline{V_{800}} \pm 1,96\frac{\sqrt{400}}{\sqrt{800}}\right] = \left[\overline{V_{800}} \pm 1,386\right]$$

2) Le théorème de Berry-Esséen assure que

$$\left|P\left(\sqrt{800}\frac{\overline{V_{800}}-a}{\sqrt{400}} \le 1,96\right) - 0,975\right| \le \frac{1}{2\sqrt{800}}\frac{E(|V-a|^3)}{(\text{Var}(W))^{3/2}} \simeq 0,0022$$

et aussi que

$$\left| P\left(\sqrt{800}\frac{\overline{V_{800}}-a}{\sqrt{400}} \leq -1,96\right) - 0,025 \right| \leq \frac{1}{2\sqrt{800}} \frac{E(|V-a|^3)}{(\mathrm{Var}(W))^{3/2}} \simeq 0,0022$$

ce qui donne $P\left(a \in I\right) \geq 0,95 - 2 \times 0,0022 \simeq 94,5\%$.

3) Dans le cas où on ne dispose que de huit mesures $V_1 = a + W_1, \ldots, V_8 = a + W_8$ le théorème central limite fournirait une approximation très mauvaise. On utilise l'inégalité de Tchebychev :

$$P\left(\left|\overline{V_8} - a\right| \geq \varepsilon\right) \leq \frac{400}{8\varepsilon^2} = \frac{50}{\varepsilon^2}$$

On choisit $\varepsilon = \sqrt{1000} \simeq 31,62$ pour avoir $\frac{50}{\varepsilon^2} = 0,05$ et on obtient

$$P\left(a \in J\right) \geq 95\% \quad \text{pour} \quad J = \left[\overline{V_8} \pm 31,62\right]$$

L'inconvénient de cet intervalle est qu'il est très large, donc fournit une estimation très imprécise.

Ex 3. 1) On peut calculer la fonction de répartition F_α comme primitive de la densité f_α soit en utilisant une intégrale soit par un calcul d'aire sur le graphe de la densité :

$$\forall t \in \mathbb{R} \quad F_\alpha(t) = (1-\alpha)(t+1)\mathbf{1}_{-1\leq t<0} + (1-\alpha+\alpha t)\mathbf{1}_{0\leq t<1} + \mathbf{1}_{t\geq 1}$$

2) La fonction quantile est définie pour $u \in]0;1[$ par $F_\alpha^{-1}(u) = \inf\{t \in \mathbb{R}\ ;\ F_\alpha(t) \geq u\}$.
Pour $u < 1-\alpha$ le calcul se ramène à $F_\alpha^{-1}(u) = \inf\{t \in [-1;0]\ ;\ (1-\alpha)(t+1) \geq u\} = \frac{u}{1-\alpha} - 1 = \frac{u-1+\alpha}{1-\alpha}$.
Pour $u > 1-\alpha$ le calcul se ramène à $F_\alpha^{-1}(u) = \inf\{t \in [0;1]\ ;\ 1-\alpha+\alpha t \geq u\} = \frac{u-1+\alpha}{\alpha}$.
Si $U \sim \mathcal{U}nif(]0;1[)$ alors $\left(\frac{U}{1-\alpha} - 1\right)\mathbf{1}_{U<1-\alpha} + \left(\frac{U-1+\alpha}{\alpha}\right)\mathbf{1}_{U>1-\alpha}$ a pour densité f_α.

3) Pour construire l'estimateur du maximum de vraisemblance, on calcule la vraisemblance en tant que fonction des observations et du paramètre inconnu. Il est plus pratique d'utiliser l'expression de la densité sous forme de produit.

$$L(x_1,\ldots,x_n,\alpha) = \prod_{i=1}^n f_\alpha(x_i) = \prod_{i=1}^n (1-\alpha)(\mathbf{1}_{-1\leq x_i\leq 1})\left(\frac{\alpha}{1-\alpha}\right)^{(\mathbf{1}_{x_i\geq 0})}$$

$$L(x_1,\ldots,x_n,\alpha) = (1-\alpha)^n(\mathbf{1}_{\forall i\leq n \quad x_i\in[-1;1]})\left(\frac{\alpha}{1-\alpha}\right)^{\sum_{i=1}^n \mathbf{1}_{x_i\geq 0}}$$

On se restreint au cas où tous les x_i sont dans $[-1;1]$ puisqu'au final on n'utilisera notre calcul que pour des tirages de densité f_α, qui appartiennent à $[-1;1]$ presque surement.

$$\ln(L(x_1,\dots,x_n,\alpha)) = n\ln(1-\alpha) + \sum_{i=1}^{n} \mathbf{1}_{x_i\geq 0}(\ln(\alpha) - \ln(1-\alpha))$$

La dérivée de ceci par rapport à α vaut

$$\frac{d}{d\alpha}\ln(L(x_1,\dots,x_n,\alpha)) = \frac{-n}{1-\alpha} + \sum_{i=1}^{n} \mathbf{1}_{x_i\geq 0}(\frac{1}{\alpha} + \frac{1}{1-\alpha}) = \frac{-n\alpha + \sum_{i=1}^{n}\mathbf{1}_{x_i\geq 0}}{\alpha(1-\alpha)}$$

donc la vraisemblance a un unique maximum, atteint en le point $\alpha = \frac{\sum_{i=1}^{n}\mathbf{1}_{x_i\geq 0}}{n}$. L'estimateur du maximum de vraisemblance du paramètre inconnu α est donc $\widehat{a_n} = \frac{\sum_{i=1}^{n}\mathbf{1}_{X_i\geq 0}}{n}$. C'est la proportion empirique de données positives.

4) Les X_i sont i.i.d. donc les $Y_i = \mathbf{1}_{X_i\geq 0}$ sont également i.i.d. et suivent la loi de Bernoulli de paramètre $P(X_i \geq 0) = \alpha$. Par conséquent l'estimateur $\widehat{a_n} = \overline{Y}_n$ a pour espérance α (il est donc sans biais) et la loi forte des grands nombres assure que $\widehat{a_n} = \overline{Y}_n$ converge presque surement vers $E(Y_1) = \alpha$ (il est donc fortement consistant). Puisqu'il est sans biais, son erreur quadratique moyenne est égale à sa variance

$$EQM(\widehat{a_n}) = \mathrm{Var}(\overline{Y}_n) = \frac{\mathrm{Var}(Y_1)}{n} = \frac{\alpha(1-\alpha)}{n}$$

5)

$$E(X_1) = \int_{-\infty}^{+\infty} x\ f_\alpha(x)\ dx = \int_{-1}^{0}(1-\alpha)x\ dx + \int_{0}^{1}\alpha x\ dx = (1-\alpha)(-\frac{1}{2}) + \alpha(\frac{1}{2}) = \alpha - \frac{1}{2}$$

$$E(X_1^2) = \int_{-\infty}^{+\infty} x^2\ f_\alpha(x)\ dx = \int_{-1}^{0}(1-\alpha)x^2\ dx + \int_{0}^{1}\alpha x^2\ dx = (1-\alpha)(\frac{1}{3}) + \alpha(\frac{1}{3}) = \frac{1}{3}$$

$$\mathrm{Var}(X_1) = \frac{1}{3} - (\alpha - \frac{1}{2})^2 = \alpha(1-\alpha) + \frac{1}{12}$$

6) Par définition $E(\overline{\alpha}_n) = E(\overline{X}_n + \frac{1}{2}) = E(X_1) + \frac{1}{2} = \alpha - \frac{1}{2} + \frac{1}{2} = \alpha$ donc $\overline{\alpha}_n$ est sans biais. Et la loi forte des grands nombres assure que $\overline{X}_n$ converge presque surement vers $E(X_1) = \alpha - \frac{1}{2}$, donc que $\overline{\alpha}_n$ converge presque surement vers α. Ce nouvel estimateur est fortement consistant. Son erreur quadratique moyenne est

$$EQM(\overline{\alpha}_n) = \mathrm{Var}(\overline{\alpha}_n) = \mathrm{Var}(\overline{X}_n + \frac{1}{2}) = \mathrm{Var}(\overline{X}_n) = \frac{\mathrm{Var}(X_1)}{n} = \frac{\alpha(1-\alpha) + \frac{1}{12}}{n}$$

C'est plus grand que l'erreur quadratique moyenne de $\hat{\alpha}_n$, donc $\hat{\alpha}_n$ est un meilleur estimateur de α que $\overline{\alpha}_n$.

Université Lille1
Sciences et Technologies

U.F.R. de Mathématiques

MASS 3 S5 Estimation **Année 2013-2014 M.Fradon**

Examen, deuxième session, 25 février 2014

Durée : 3 heures.

Matériel autorisé : calculatrice, tables statistiques, feuille manuscrite A4 recto-verso.

Ex 1. X est une variable aléatoire de fonction de répartition F :

$$\forall t \in \mathbb{R} \quad F(t) = 2^t \, \mathbf{1}_{t<0} + \mathbf{1}_{t\geq 0}$$

1) Déterminer sa fonction quantile.

2) Calculer l'espérance et la variance de X.

3) On effectue 30 tirages indépendants selon la loi de X. Montrer qu'il y a moins d'une chance sur cent que leur moyenne empirique soit inférieure à -2.

4) Indiquer comment faire un tirage de X à l'aide de `scilab`.

Ex 2. L'estimation d'une proportion inconnue nécessite des techniques statistique, mais aussi des données fiables auxquelles les appliquer. Or certaines données sont difficilement accessibles. Comment estimer la proportion de gens qui volent au supermarché ? Qui se lavent moins d'une fois par semaine ? Qui fraudent le fisc ? A ce type de question, il y a une « bonne » réponse, socialement acceptable. Le sondé est tenté de mentir, par crainte de la réaction du sondeur. Cet exercice présente une technique servant à recueillir des données fiables dans un tel cas.

Le sondeur fournit un paquet d'une centaine de cartes. Il laisse au sondé le temps de le feuilleter : environ la moitié des cartes portent l'inscription « *oui=rouge, non=jaune* », les autres cartes portent l'inscription « *oui=jaune, non=rouge* ». A chaque question, le sondé tire une carte au hasard. Il répond « jaune » ou « rouge » selon le code inscrit sur la carte. Il remet la carte dans le paquet sans l'avoir montrée au sondeur.

Avec ce système, il n'y a plus de réponse gênante. Le sondé n'est pas tenté de mentir, ce qui contribue à la fiabilité des données.

On note c la proportion de cartes « *oui=rouge, non=jaune* ». Pour inspirer confiance aux sondés, il faut que c soit proche de 50%. On veille cependant à ce que c ne soit pas exactement égal à $0,5$.

On note p la proportion de gens ayant déjà volé au supermarché. p est inconnue et on souhaite l'estimer.

1) On choisit une personne au hasard. On lui pose la question « Avez-vous déjà volé dans un supermarché ? ». Elle tire une carte et répond selon la procédure décrite. Exprimer en fonction de p et c la probabilité r que la personne réponde « rouge ».

2) On pose cette même question à $n = 625$ personnes choisies au hasard. On note X_i la variable aléatoire qui vaut 1 si le $i^{\text{ème}}$ sondé répond « rouge », et qui vaut 0 sinon. Construire un intervalle de confiance pour l'estimation de r, au niveau de confiance 90%.

3) Pour ce sondage, la proportion de cartes « *oui=rouge, non=jaune* »est $c = 0,55$. En déduire un intervalle de confiance pour l'estimation de p, toujours au niveau de confiance 90%. Expliquer pourquoi on constitue toujours le paquet de cartes avec $c \neq 0,5$ (quel problème poserait un c égal à 50% ?).

4) A partir des X_i, construire l'estimateur du maximum de vraisemblance de r. En déduire un estimateur de la proportion inconnue p.

Ex 3. On a l'habitude de calculer l'espérance et la variance d'une loi à partir des paramètres de cette loi. Inversement, on peut aussi exprimer le ou les paramètres à partir de l'espérance et de la variance. On veut utiliser cette idée pour construire des estimateurs fortement consistants des paramètres d'une loi.

On note X une variable aléatoire dont la loi P_θ dépend d'un paramètre θ. La valeur du paramètre est inconnue. On connait seulement l'ensemble Θ de tous les paramètres possibles. X a une espérance qui vaut $E_\theta(X)$ quand le paramètre vaut θ.

1) On effectue n tirages indépendants selon la loi P_θ. Quel résultat fournit un estimateur fortement consistant de $E_\theta(X)$? Le rappeler.

2) On dispose d'une fonction continue g telle que $g(E_\theta(X)) = \theta$ pour tout $\theta \in \Theta$. Construire à partir de g un estimateur fortement consistant du paramètre inconnu θ.

3) Dans le cas où X suit une loi exponentielle de paramètre inconnu, quel est l'ensemble Θ ? Quelle est la fonction g ? Quel estimateur obtient-on pour le paramètre inconnu ?

Université Lille1 Sciences et Technologies

U.F.R. de Mathématiques

MASS 3 S5 Estimation **Année 2013-2014 M.Fradon**

Corrigé de l'examen de deuxième session, février 2014

Ex 1. 1) Puisque $F(t) = 2^t \ \mathbf{1}_{t<0} + \mathbf{1}_{t\geq 0}$ pour tout $t \in \mathbb{R}$

$$\forall u \in]0;1[\quad F^{-1}(u) = \inf\{t \in \mathbb{R}\,;\, F(t) \geq u\} = \inf\{t \in \mathbb{R}\,;\, 2^t \geq u\} = \inf\{t \in \mathbb{R}\,;\, t\ln(2) \geq \ln(u)\} = \frac{\ln(u)}{\ln(2)}$$

2) Remarquons que $P(X \leq 0) = F(0) = 0$ donc $P(|X| > t) = P(X < -t) = 2^{-t}$ pour t positif.

$$E(|X|) = \int_0^{+\infty} P(|X| > t)\ dt = \int_0^{+\infty} e^{-t\ln(2)}\ dt = \left[\frac{e^{-t\ln(2)}}{-\ln(2)}\right]_0^{+\infty} = \frac{1}{\ln(2)} < +\infty$$

donc X a une espérance et $E(X) = E(-|X|) = \frac{-1}{\ln(2)}$.

Pour la variance, on utilise une densité f de X, qu'on obtient en dérivant F puisque F est continue $\mathcal{C}^1$ par morceaux.

$$\forall x \in \mathbb{R} \quad f(x) = \ln(2)2^x \mathbf{1}_{x<0} \quad \text{donc} \quad E(X^2) = \int_{-\infty}^{+\infty} x^2\ f(x)\ dx = \int_{-\infty}^{0} x^2 \ln(2)\ 2^x\ dx$$

$$E(X^2) = \left[x^2\ 2^x\right]_{-\infty}^{0} - 2\int_{-\infty}^{0} x\ 2^x\ dx = -\frac{2}{\ln(2)}\int_{-\infty}^{+\infty} x\ f(x)\ dx = -\frac{2}{\ln(2)}E(X) = \frac{2}{(\ln(2))^2}$$

Finalement, $\mathrm{Var}(X) = \frac{2}{(\ln(2))^2} - \frac{1}{(\ln(2))^2} = \frac{1}{(\ln(2))^2}$.

3) Notons $\overline{X}_{30}$ la moyenne empirique des 30 tirages indépendants. Comme $\frac{1}{\ln(2)} \simeq 1,44$

$$P\left(\overline{X}_{30} \leq -2\right) \leq P\left(\overline{X}_{30} + \frac{1}{\ln(2)} \leq -0,5\right) \leq P\left(\left|\overline{X}_{30} + \frac{1}{\ln(2)}\right| \geq 0,5\right)$$

L'inegalité de Tchebychev donne alors

$$P\left(\overline{X}_{30} \leq -2\right) \leq \frac{1}{(\ln(2))^2\ 30^2\ (0,5)^2} < 0,01$$

4) On peut simuler X par inversion de la fonction de répartition : $\frac{\ln(U)}{\ln(2)}$ a même loi que X si U suit la loi uniforme sur $]0;1[$. Il suffit donc de faire `log(rand(1,'uniform'))/log(2)` pour effectuer un tirage de X à l'aide de `scilab`.

Ex 2. Avez-vous déjà volé dans un supermarché ?

1) Définissons les événements
$V = \{$la personne a déjà volé dans un supermarché$\}$
$C = \{$la carte porte l'inscription "*oui=rouge, non=jaune*"$\}$
$R = \{$la personne répond "rouge"$\}$
$R = (V \cap C) \cup (V^c \cap C^c)$ et V et C sont indépendants donc

$$r = P(R) = P(V\cap C)+P(V^c\cap C^c) = P(V)P(C)+P(V^c)P(C^c) = pc+(1-p)(1-c) = 1-c+p(2c-1)$$

2) Les X_i sont i.i.d. de loi de $\mathcal{B}er(r)$. On a $n = 625$ données, on va utiliser le théorème central limite, avec autonormalisation car la variance $r(1-r)$ des X_i est inconnue :

$$P\left(\left|\sqrt{n}\frac{\overline{X_n} - r}{\sqrt{\overline{X_n}(1-\overline{X_n})}}\right| \leq 1,645\right) \simeq 90\%$$

c'est-à-dire $P\left(U_1 \leq r \leq U_2\right) \simeq 90\%$ où $U_1 = \overline{X_n} - 1,645\frac{\sqrt{\overline{X_n}(1-\overline{X_n})}}{\sqrt{n}}$ et $U_2 = \overline{X_n} + 1,645\frac{\sqrt{\overline{X_n}(1-\overline{X_n})}}{\sqrt{n}}$.

3) On sait que $r = 1 - c + p(2c-1)$. L'intervalle de confiance précédent donne $P\left(U_1 \leq 1 - c + p(2c-1) \leq U_2\right) \simeq 90\%$ i.e.

$$P\left(\frac{U_1 - 1 + c}{2c-1} \leq p \leq \frac{U_1 - 1 + c}{2c-1}\right) \simeq 90\%$$

Puisque $c = 0,55$ ici, $P(p \in I) \simeq 90\%$ où

$$I = \left[10\left(\overline{X_n} - 1,645\frac{\sqrt{\overline{X_n}(1-\overline{X_n})}}{\sqrt{n}} - 0,45\right)\ ;\ 10\left(\overline{X_n} + 1,645\frac{\sqrt{\overline{X_n}(1-\overline{X_n})}}{\sqrt{n}} - 0,45\right)\right]$$

On ne peut évidemment construire cet intervalle de confiance que parce que $2c - 1 \neq 0$ i.e. $c \neq 0,5$ ici. Si le jeu de cartes est tel que $c = 0,5$, la réponse du sondé ("rouge" ou "jaune") devient indépendante du message à transmettre ("oui" ou "non") et cette technique ne permet plus d'évaluer la proportion de réponses "oui" à partir de la proportion de réponses "rouge".

4) On doit construire l'estimateur du maximum de vraisemblance de r. On calcule la vraisemblance

$$L(x_1, \ldots, x_n, r) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} r^{x_i}(1-r)^{1-x_i} = r^{\sum_{i=1}^n x_i}(1-r)^{n-\sum_{i=1}^n x_i}$$

On passe au logarithme en utilisant le fait que $0 < r < 1$ (il existe des gens qui volent, et il en existe qui ne volent pas)

$$\ln L(x_1, \ldots, x_n, \theta) = \sum_{i=1}^{n} x_i \ln(r) + (n - \sum_{i=1}^{n} x_i)\ln(1-r)$$

On dérive

$$\frac{d}{dr}\ln L(x_1, \ldots, x_n, r) = \frac{1}{r}\sum_{i=1}^{n} x_i - \frac{1}{1-r}(n - \sum_{i=1}^{n} x_i) = \frac{1}{r(1-r)}\left(-nr + \sum_{i=1}^{n} x_i\right)$$

et on trouve que la vraisemblance est maximale en $r = \frac{1}{n}\sum_{i=1}^n x_i$. Donc l'estimateur du maximum de vraisemblance est $\widehat{r_n} = \overline{X_n}$.

Puisque $p = \frac{r-1+c}{2c-1} = 10(r - 0,45)$ on peut estimer la proportion inconnue p par $\widehat{p_n} = 10(\overline{X_n} - 0,45)$.

Ex 3. Une machine à estimateurs : la méthode des moments

1) La loi forte des grands nombres, appliquée à la loi intégrable P_θ, assure que la moyenne empirique $\overline{X_n}$ de n tirages indépendants de loi P_θ converge presque surement vers l'espérance $E_\theta(X)$. Autrement dit $\overline{X_n}$ est un estimateur fortement consistant de $E_\theta(X)$.

2) La convergence presque sûre se conserve par composition avec une fonction continue, donc

$$\overline{X_n} \xrightarrow[n\to+\infty]{p.s.} E_\theta(X) \quad \text{implique} \quad g(\overline{X_n}) \xrightarrow[n\to+\infty]{p.s.} g(E_\theta(X)) = \theta$$

$g(\overline{X_n})$ est donc un estimateur fortement consistant de θ.

3) Dans le cas où X suit la loi $\mathcal{E}xp(a)$ avec $a > 0$ inconnu, l'ensemble des valeurs possibles pour le paramètre $\theta = a$ est $\Theta =]0; +\infty[$. Quand le paramètre vaut a l'espérance de X vaut $1/a$ i.e. $E_a(X) = 1/a$. Il suffit de choisir pour g la fonction définie sur $\Theta =]0; +\infty[$ par $g(x) = 1/x$ et on a $g(E_a(X)) = a$. On obtient alors l'estimateur

$$g(\overline{X_n}) = \frac{1}{\overline{X_n}} \xrightarrow[n\to+\infty]{p.s.} g(E_a(X)) = a$$

Université Lille1 Sciences et Technologies

U.F.R. de Mathématiques

MASS 3 S5 Estimation **Année 2013-2014 M.Fradon**

Devoir surveillé, 24 octobre 2013

Durée : 2 heures.

Matériel autorisé : calculatrice, tables statistiques, feuille manuscrite A4 recto-verso.

Ex 1. 56 patients souffrant de douleurs dorsales sont volontaires pour tester une thérapie nouvelle, qu'on envisage d'étendre ensuite à toute la population. On cherche à évaluer l'efficacité du traitement, c'est-à-dire la proportion p de personnes que cette technique peut soulager, parmi toute la population souffrant de mal de dos.

1) La thérapie comporte deux séances. On note X le nombre de patients qui déclarent, un mois après la deuxième séance, avoir constaté une disparition ou une réduction durable des douleurs. En utilisant l'inégalité de Tchebychev, construire à partir de X un intervalle de confiance au niveau 90% pour l'estimation de p.

2) L'enquète menée un mois après la deuxième séance donne le résultat suivant : parmi les 56 patients volontaires, 21 ont constaté un effet positif. Quelle est la valeur observée de l'intervalle de confiance construit ?

Ex 2. *Les deux parties de cet exercice sont indépendantes.*

On note F la fonction de répartition suivante :

$$\forall t \in \mathbb{R} \quad F(t) = \frac{t^3+4}{8}\mathbf{1}_{[-1;1[}(t) + \mathbf{1}_{[1;+\infty[}(t)$$

Partie 1

1) Tracer le graphe de F. La loi correspondante est-elle à densité ? (justifier).

2) Déterminer la fonction quantile. Construire à partir d'un tirage uniforme sur $]0;1[$ une variable aléatoire de fonction de répartition F.

Partie 2

3) $U_1, U_2, U_3, \ldots$ sont des variables aléatoires indépendantes de loi uniforme sur $]0;1[$. $V_1, V_2, V_3, \ldots$ sont des variables aléatoires indépendantes entre elles et indépendantes des U_i telles que $P(V_i = 1) = P(V_i = -1) = \frac{1}{2}$ pour tout i.
On fabrique une suite de variables aléatoires en posant pour chaque i de $\mathbb{N}^*$

$$W_i = V_i \min(1; \sqrt[3]{4U_i})$$

Prouver que W_1 a pour fonction de répartition F.

4) Montrer que quand n tend vers l'infini, la moyenne empirique $\overline{W_n} = \sum_{i=1}^n W_i/n$ converge presque surement vers une limite que l'on déterminera.

5) En admettant que W_1 a pour variance $\frac{9}{10}$ calculer une valeur approximative de la probabilité de l'évènement $\{\sum_{i=1}^{1000} W_i \leq \frac{1}{100}\}$.

6) On admet également que $E(|W_1|^3) = \frac{7}{8}$. En déduire la précision du calcul approximatif précédent, et donner un encadrement de $P\left(\sum_{i=1}^{1000} W_i \leq \frac{1}{100}\right)$.

7) Prouver ce qui a été admis auparavant : calculer $\mathrm{Var}(W_1)$ et $E(|W_1|^3)$.

Université Lille1 Sciences et Technologies

U.F.R. de Mathématiques

MASS 3 S5 Estimation **Année 2013-2014 M.Fradon**

Corrigé du devoir surveillé du 24 octobre 2013

Ex 1. 1) On suppose que les personne souffrant du dos sont indépendantes et que chaque personne choisie au hasard a une probablité p d'être soulagée par la thérapie nouvelle. Le nombre X de patients soulagés à l'issue de l'expérience suit alors la loi $\mathcal{B}in(56;p)$. L'inégalité de Tchebychev s'écrit $P(|X-56p|\geq t)\leq \frac{56p(1-p)}{t^2}$ pour chaque t strictement positif.
La majoration de $p(1-p)$ par $\frac{1}{4}$ donne $P(|X-56p|\geq t)\leq \frac{56}{4t^2}$ i.e. $P(|X-56p|\geq t)\leq \frac{14}{t^2}$. On choisit t pour avoir $\frac{14}{t^2}=0,1$: $t=\sqrt{140}$.

$$P\left(\left|\frac{X}{56}-p\right|\geq\frac{\sqrt{140}}{56}\right)\leq 0,1 \quad\Longleftrightarrow\quad P\left(\left|\frac{X}{56}-p\right|<\frac{\sqrt{140}}{56}\right)\geq 0,9$$

donc $I=\left]\frac{X}{56}-\frac{\sqrt{140}}{56};\frac{X}{56}+\frac{\sqrt{140}}{56}\right[$ est un intervalle au niveau de confiance 90% pour l'estimation de p.

2) Après avoir réalisé les séances, on recontacte les patients pour s'enquérir de leur état de santé et on obtient pour ce tirage $X(\omega)=21$ donc

$$I(\omega)=\left]\frac{21}{56}-\frac{\sqrt{140}}{56};\frac{21}{56}+\frac{\sqrt{140}}{56}\right[\simeq]0,3750\pm 0,2113[=]16,37\%;58,63\%[$$

Cet intervalle est large, comme on pouvait s'y attendre au vu du faible nombre de patients participant à l'expérimentation.

Ex 2.

Partie 1

1) F est constante à gauche et à droite de $[-1;1[$. Sur cet intervalle elle croit de 3/8 à 5/8. Attention, le graphe n'est pas une droite sur $[-1;1[$. On voit sans calcul que l'allure est celle d'un polynôme de degré trois avec pente nulle en zéro et pente 3/8 en -1 et 1.

Le graphe de F comporte des sauts : $F(-1^-)=0<F(-1)=\frac{3}{8}$ et $F(1^-)=\frac{5}{8}<F(1)=1$. La fonction de répartition F est discontinue donc la loi n'est pas à densité.

2) Rappel : les quantiles pour u hors de $]0;1[$ n'existent pas puisque les probabilités à valeurs hors de $[0;1]$ n'existent pas.
Pour u dans $]0;1[$, la fonction quantile en u vaut $F^{-1}(u)=\inf\{t\in\mathbb{R};F(t)\geq u\}$.

Dans le cas où $u \leq \frac{3}{8}$, par définition $\{t \in \mathbb{R}; F(t) \geq u\} = [-1; +\infty[$ et donc $F^{-1}(u) = -1$.
Dans le cas où $\frac{3}{8} < u < \frac{5}{8}$,

$$F(t) \geq u \quad \Longleftrightarrow \quad \Big(t \geq 1 \ \text{ ou } \ -1 \leq t < 1 \text{ et } t^3 + 4 \geq 8u\Big) \quad \Longleftrightarrow \quad t^3 \geq 8u - 4$$

donc $F^{-1}(u) = \sqrt[3]{8u-4}$.
Dans le cas où $u \geq \frac{5}{8}$, on a $\{t \in \mathbb{R}; F(t) \geq u\} = [1; +\infty[$ donc $F^{-1}(u) = 1$. En résumé :

$$\forall u \in]0;1[\quad F^{-1}(u) = -\mathbf{1}_{u \leq \frac{3}{8}} + \sqrt[3]{8u-4}\ \mathbf{1}_{\frac{3}{8}<u<\frac{5}{8}} + \mathbf{1}_{u \geq \frac{5}{8}}$$

et donc $-\mathbf{1}_{U \leq \frac{3}{8}} + \sqrt[3]{8U-4}\ \mathbf{1}_{\frac{3}{8}<U<\frac{5}{8}} + \mathbf{1}_{U \geq \frac{5}{8}}$ a pour fonction de répartition F si U suit la loi uniforme sur $]0;1[$.

Partie 2

3) Fixons un réel t et calculons $F_{W_1}(t)$.

$$P(W_1 \leq t) = P(V_1 \min(1; \sqrt[3]{4U_1}) \leq t \mid V_1 = 1) \times \frac{1}{2} + P(V_1 \min(1; \sqrt[3]{4U_1}) \leq t \mid V_1 = -1) \times \frac{1}{2}$$

L'indépendance entre U_1 et V_1 assure que

$$F_{W_1}(t) = \frac{1}{2} P(\min(1; \sqrt[3]{4U_1}) \leq t) + \frac{1}{2} P(\min(1; \sqrt[3]{4U_1}) \geq -t)$$

La première de ces deux probabilités est nulle si t est négatif, égale à 1 si $t \geq 1$ et entre les deux (pour $0 \leq t \leq 1$) elle vaut $P(\sqrt[3]{4U_1} \leq t) = P(U_1 \leq t^3/4) = \frac{t^3}{4}$. La deuxième est nulle si $-t \geq 1$ i.e. $t \leq -1$, égale à 1 si t est positif, et entre les deux (pour $-1 \leq t \leq 0$) elle vaut $P(\sqrt[3]{4U_1} \geq -t) = P(U_1 \geq -t^3/4) = 1 - (-\frac{t^3}{4})$. D'où

$$F_{W_1}(t) = \begin{cases} 0 & \text{si } t \leq -1 \\ \frac{1}{2} \times 0 + \frac{1}{2} \times (1 + \frac{t^3}{4}) & \text{si } -1 < t \leq 0 \\ \frac{1}{2} \times \frac{t^3}{4} + \frac{1}{2} \times 1 & \text{si } 0 < t \leq 1 \\ \frac{1}{2} \times 1 + \frac{1}{2} \times 1 & \text{si } t \geq 1 \end{cases}$$

La fonction de répartition de W_1 est égale à F.

4) $W_1 = V_1 \min(1; \sqrt[3]{4U_1})$ est bornée (à valeurs dans $[-1;1]$) donc elle a des moments de tous ordres et en particulier une espérance. L'indépendance entre V_1 et U_1, et la nullité de l'expérance de V_1 donnent

$$E(W_1) = E(V_1) E(\min(1; \sqrt[3]{4U_1})) = 0 \times E(\min(1; \sqrt[3]{4U_1})) = 0$$

La loi forte des grands nombres, appliquée aux variables intégrables i.i.d. d'espérance nulle W_i, assure la convergence de la moyenne empirique vers zéro

$$\overline{W_n} \xrightarrow[n \to +\infty]{p.s.} 0$$

5) Les W_i sont i.i.d. de variance $0,9 > 0$ et d'espérance nulle, donc d'après le théorème central limite $\frac{\sum_{i=1}^n W_i}{\sqrt{0,9n}}$ converge en loi vers $Z \sim \mathcal{N}(0;1)$. Pour $n = 1000$

$$P\left(\sum_{i=1}^{1000} W_i \leq \frac{1}{100}\right) = P\left(\frac{\sum_{i=1}^{1000} W_i}{\sqrt{900}} \leq \frac{1}{100\sqrt{900}}\right) = P\left(\frac{\sum_{i=1}^{1000} W_i}{\sqrt{900}} \leq \frac{1}{3000}\right)$$

est proche de

$$P\left(Z \leq \frac{1}{3000}\right) \simeq 0,5$$

6) En admettant que $E(|W_1|^3) = \frac{7}{8}$ le théorème de Berry-Esséen permet d'affirmer que

$$\left|P\left(\sum_{i=1}^{1000} W_i \leq \frac{1}{100}\right) - P\left(Z \leq \frac{1}{3000}\right)\right| \leq \frac{0,5 \times \frac{7}{8}}{\sqrt{1000}(0,9)^{3/2}}$$

i.e.

$$\left|P\left(\sum_{i=1}^{1000} W_i \leq \frac{1}{100}\right) - 0,5\right| \leq 0,016204$$

On en conclut que $P\left(\sum_{i=1}^{1000} W_i \leq \frac{1}{100}\right)$ est entre $0,5 - 0,016204 \simeq 48,38\%$ et $0,5 + 0,016204 \simeq 51,62\%$.

7) Il reste à calculer la variance et le moment centré d'ordre trois de W_1. L'espérance de W_1 est nulle donc $\mathrm{Var}(W_1) = E(W_1^2)$. Par indépendance, et en remarquant que $|V_1| = 1$ presque surement,

$$E(W_1^2) = E(V_1^2)E(\min(1; (4U_1)^{2/3}) = 1 \times \int_0^{+\infty} P(\min(1; (4U_1)^{2/3}) > t)\ dt$$

Pour que leur minimum soit supérieur à t il est nécessaire et suffisant que tous les deux soient supérieur à t

$$P(\min(1; (4U_1)^{2/3}) > t) = P(1 > t \text{ et } (4U_1)^{2/3}) > t) = \mathbf{1}_{1>t} P((4U_1)^{2/3} > t) = \mathbf{1}_{1>t} P(4U_1 > t^{3/2})$$

et la fonction de répartition U_1, qui suit une loi uniforme, donne le résultat

$$\mathrm{Var}(W_1) = \int_0^1 P(U_1 > \frac{t^{3/2}}{4})\ dt = \int_0^1 1 - \frac{t^{3/2}}{4}\ dt = 1 - \frac{1}{4}\left[\frac{t^{5/2}}{5/2}\right]_0^1 = 1 - \frac{1}{4} \times \frac{2}{5} = 0,9$$

On peut procéder de façon analogue pour le calcul de $E(|W_1|^3)$ qui est plus simple :

$$E(|W_1|^3) = 1 \times E(\min(1; 4U_1)) = \int_0^{+\infty} P(\min(1; 4U_1) > t)dt = \int_0^1 P(4U_1 > t)dt$$

$$E(|W_1|^3) = \int_0^1 1 - \frac{t}{4}\ dt = 1 - \frac{1}{8} = \frac{7}{8}$$

Dans certaines copies, on voit des connaissances de base (fonction de répartition, calcul d'espérance) moins bien maitrisées que des techniques plus avancées. L'utilisation de la loi forte des grands nombres, du théorème central limite et du théorème de Berry-Esséen n'est pas possible sans calcul d'espérance et variance. Il n'est pas trop tard pour s'entrainer à calculer des espérances ou des lois, ça peut être fait sur des exercices de Semestre 5, et ça peut rapporter gros pour ceux qui ont perdu des points dans la partie 2 de l'exercice 2.