Université Lille1 Sciences et Technologies

U.F.R. de Mathématiques

MASS 3 S5 Estimation **Année 2012-2013 M.Fradon**

Examen, 19 décembre 2012

Durée : 3 heures.

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

Ex 1. En début d'année, on décide d'observer 22 parcelles de blé tendre. On notera $X_1,\dots,X_{22}$ les rendements (en quintaux par hectare) que vont donner ces parcelles. Les X_i sont aléatoires, indépendants et de même loi. On les suppose gaussiens. On souhaite évaluer le rendement moyen $m = E(X_i)$ et l'écart-type $\sigma = \sqrt{\mathrm{Var}(X_i)}$.

1) Construire à partir des X_i un intervalle de confiance I pour l'estimation du rendement moyen $m = E(X_1)$ au niveau de confiance 99%.

2) Construire également à partir des X_i un intervalle de confiance J pour l'estimation de l'écart-type $\sigma = \sqrt{\mathrm{Var}(X_1)}$ des rendements, toujours au niveau de confiance 99%.

3) Calculer les valeurs observées de I et J qu'on obtient quand, en fin d'année, on constate que $\sum_{i=1}^{22} X_i(\omega) = 1335 q/ha$ et $\sum_{i=1}^{22} X_i^2(\omega) = 84976 (q/ha)^2$.

Ex 2. La loi de Burr est utilisée pour modéliser les revenus des ménages. Cette loi a deux paramètres réels $c > 0$ et $a > 0$. Dans le cadre de cet exercice, c est fixé et connu. On cherche à estimer a.

La densité de la loi de Burr de paramètres c et a est : $\forall x \in \mathbb{R} \quad f_a(x) = \dfrac{a\ c\ x^{c-1}}{(1+x^c)^{a+1}} \mathbf{1}_{x>0}$.

1) A partir des revenus $X_1, X_2, \dots, X_n$ de n ménages choisis au hasard, construire l'estimateur du maximum de vraisemblance $\widehat{a_n}$ du paramètre inconnu a.

2) Déterminer la fonction de répartition F_a et la fonction quantile F_a^{-1} de la loi de Burr.

Ex 3. On veut simuler la loi de densité $f_X : x \mapsto \dfrac{8}{\pi} (\cos(x)\sin(x))^2 \mathbf{1}_{0<x<\pi}$.

1) Déterminer une densité de probabilité f_W et une constante C telles que $f_X \leq C\ f_W$.

2) Indiquer précisément la démarche à suivre pour faire un tirage selon la loi demandée.

Ex 4. Chaque année, l'OFIP [1] réalise une enquête sur la situation des diplômés de Lille 1. Deux ans après l'obtention de leur Master Professionnel, d'anciens étudiants sont interrogés : on leur demande s'il ont un emploi, dans quel secteur, etc.

1. Observatoire des Formations et de l'Insertion Professionnelle : organisme de Lille 1 qui réalise et publie des statistiques sur les formations et les étudiants.

1) On s'intéresse d'abord aux masters du secteur Economie-Gestion-Management-Marketing. Sur les 526 diplômés interrogés, 487 ont un emploi. En indiquant quelle(s) variable(s) aléatoire(s) et quel(s) théorème(s) vous utilisez, construire un intervalle de confiance au niveau 95% pour l'estimation de la proportion p_E de diplômés qui ont un emploi.

2) Une enquête similaire a été menée sur les masters du secteur Biologie-Géologie. Sur les 131 diplômés interrogés, 110 ont un emploi. Quel est l'intervalle de confiance à 95% pour l'estimation de la proportion p_B de diplômés de Biologie qui ont un emploi ? Peut-on raisonnablement affirmer que les taux d'insertion professionnelle p_E et p_B sont différents ?

Ex 5. Au bord d'une route, un dispositif enregistre les instants de passage des véhicules. La durée aléatoire T_i (en minutes) entre le passage d'un véhicule et le passage du véhicule suivant suit la loi exponentielle de paramètre 1. On observe ainsi les durées T_1 (après le passage du premier véhicule), T_2 (après le passage du deuxième), T_3 (après le passage du troisième), etc. Ces durées sont indépendantes.

1) Indiquer (sans calcul) l'espérance et la variance des T_i.

2) Par quelle loi peut-on approcher la durée totale $\sum_{i=1}^{240} T_i$ qui s'écoule entre le premier et le $241^{\text{ème}}$ véhicule ?

3) Quelle est approximativement la probabilité que cette durée totale dépasse 4 heures et demie ?

4) Le calcul de probabilité effectué à la question précédente est approximatif. En utilisant le théorème de Berry-Esséen, déterminer la précision de ce calcul.

Université Lille1 Sciences et Technologies

U.F.R. de Mathématiques

MASS 3 S5 Estimation **Année 2012-2013 M.Fradon**

Corrigé de l'examen du 19 décembre 2012

Ex 1. Les 22 rendements $X_1,\dots,X_{22}$ sont indépendants et de même loi gaussienne dont on peut supposer que sa variance est non-nulle. On note $\overline{X_{22}}$ et V_{22} leur espérance empirique et variance empirique. Le théorème de Student assure que

$$\sqrt{21}\frac{\overline{X_{22}}-m}{\sqrt{V_{22}}} \sim \text{Student}(21) \quad \text{et} \quad \frac{22\ V_{22}}{\sigma^2} \sim \chi^2(21)$$

1) La table de la Student(21) et la symétrie de cette loi donnent :

$$P\left(\left|\sqrt{21}\frac{\overline{X_{22}}-m}{\sqrt{V_{22}}}\right| \leq 2,831\right) \simeq 99\%$$

d'où l'intervalle de confiance sur $m = E(X_1)$: $I = \left[\overline{X_{22}} \pm 2,831\frac{\sqrt{V_{22}}}{\sqrt{21}}\right]$

2) On utilise la table du $\chi^2(21)$ qui donne :

$$99\% \simeq P\left(8,034 \leq \frac{22\ V_{22}}{\sigma^2} \leq 41,401\right) = P\left(\frac{22\ V_{22}}{8,034} \geq \sigma^2 \geq \frac{22\ V_{22}}{41,401}\right)$$

d'où l'intervalle de confiance sur $\sigma = \sqrt{\text{Var}(X_1)}$: $J = \left[\sqrt{\frac{22\ V_{22}}{41,401}}\ ;\ \sqrt{\frac{22\ V_{22}}{8,034}}\right]$.

3) $\overline{X_{22}}(\omega) = \frac{1335}{22} \simeq 60,68$ et $V_{22}(\omega) = \frac{84976}{22} - (\overline{X_{22}}(\omega))^2 \simeq 180,26$ donc

$$I(\omega) \simeq [52,38\ ;\ 68,98] \quad \text{et} \quad J(\omega) \simeq [9,78\ ;\ 22,22]$$

Ex 2. 1) On calcule d'abord la vraisemblance, qui est une fonction des observations et du paramètre inconnu a :

$$L(x_1,\dots,x_n,a) = \prod_{i=1}^n f_a(x_i) = \prod_{i=1}^n \frac{a\ c\ x_i^{c-1}}{(1+x_i^c)^{a+1}}\mathbf{1}_{x_i>0} = a^n c^n \frac{(x_1\cdots x_n)^{c-1}}{(\prod_{i=1}^n(1+x_i^c))^{a+1}}\mathbf{1}_{\min_{1\leq i\leq n} x_i\ >0}$$

Il est équivalent de maximiser la vraisemblance ou la log-vraisemblance. On se restreint au cas où tous les x_i sont strictement positifs pour définir celle-ci :

$$\ln(L(x_1,\dots,x_n,a)) = n\ln(a) + n\ln(c) + (c-1)\ln(x_1\cdots x_n) - (a+1)\sum_{i=1}^n \ln(1+x_i^c)$$

Sa dérivée par rapport à a vaut : $\frac{d}{da}\ln(L(x_1,\dots,x_n,a)) = \frac{n}{a} - \sum_{i=1}^n \ln(1+x_i^c)$.
La vraisemblance admet un unique maximum, en le point $a = \frac{n}{\sum_{i=1}^n \ln(1+x_i^c)}$. L'estimateur du maximum de vraisemblance du paramètre inconnu a est donc $\widehat{a_n} = \frac{n}{\sum_{i=1}^n \ln(1+X_i^c)}$.

2) Pour tout réel t :

$$F_a(t) = \int_{-\infty}^{t} \frac{a\ c\ x^{c-1}}{(1+x^c)^{a+1}} \mathbf{1}_{x>0}\ dx = \mathbf{1}_{t>0}\ \int_0^t a\ c\ x^{c-1}\ (1+x^c)^{-a-1}\ dx$$

On utilise que $ku'u^{k-1}$ est la dérivée de u^k (on peut aussi utiliser le changement de variable $y = x^c$ si on préfère) ce qui donne :

$$F_a(t) = \mathbf{1}_{t>0}\ \left[-(1+x^c)^{-a}\right]_0^t = \mathbf{1}_{t>0}\ (1-(1+t^c)^{-a})$$

On calcule ensuite la fonction quantile correspondante :

$$\forall u \in]0;1[\quad F_a^{-1}(u) = \inf\{t \in \mathbb{R}\ ;\ F_a(t) \geq u\} = \inf\{t > 0\ ;\ 1-(1+t^c)^{-a} \geq u\}$$

$$1-(1+t^c)^{-a} \geq u \quad \Leftrightarrow \quad \frac{1}{(1+t^c)^a} \leq 1-u \quad \Leftrightarrow \quad (1+t^c)^a \geq \frac{1}{1-u} \quad \Leftrightarrow \quad 1+t^c \geq \frac{1}{(1-u)^{1/a}}$$

$$\Leftrightarrow \quad t^c \geq \frac{1}{(1-u)^{1/a}} - 1 \quad \Leftrightarrow \quad t \geq ((1-u)^{-1/a}-1)^{1/c}$$

donc $F_a^{-1}(u) = ((1-u)^{-1/a}-1)^{1/c}$.

Ex 3. 1) La densité $f_X : x \mapsto \frac{8}{\pi}\ (\cos(x)\sin(x))^2\ \mathbf{1}_{0<x<\pi}$ est nulle en dehors de l'intervalle $]0;\pi[$ et majorée par $\frac{8}{\pi}$ sur cet intervalle. On peut donc choisir pour f_W la densité de la loi uniforme sur $]0;\pi[$, qui vaut $\frac{1}{\pi}\mathbf{1}_{]0;\pi[}$. En prenant $C = 8$, on aura bien $f_X \leq C\ f_W$.

2) On tire d'abord un nombre W_1 au hasard (uniformément) entre 0 et π, puis on tire un U_1 au hasard entre 0 et 1 (uniformément, et indépendamment de W_1). La condition d'acceptation $C\ f_W(W_1)\ U_1 \leq f_X(W_1)$ s'écrit ici

$$8\frac{1}{\pi}\mathbf{1}_{]0;\pi[}(W_1)U_1 \leq \frac{8}{\pi}\ (\cos(W_1)\sin(W_1))^2\ \mathbf{1}_{0<W_1<\pi} \quad \text{i.e.} \quad U_1 \leq (\cos(W_1)\sin(W_1))^2$$

Si elle est satisfaite on accepte W_1 sinon on le rejette et on recommence la procédure avec un W_2 et un U_2, éventuellement un W_3 et un U_3, etc jusqu'à ce qu'on accepte un W_i. La loi de cette valeur acceptée est bien celle de densité f_X.

Ex 4. Taux d'insertion professionnelle [2]

1) On prévoit d'interroger 526 diplômés de masters du secteur Economie-Gestion-Management-Marketing. On notera X_i l'indicatrice du fait que l'ex-étudiant i a un emploi. On suppose les X_i indépendants, et ils suivent tous la loi de Bernoulli de paramètre p_E. Ce paramètre ne vaut ni 0 ni 1 puisqu'il y a effectivement des diplômés en emploi et des diplômés au chômage. La proportion empirique $\overline{X_n}$ de personnes interrogées ayant un emploi a une loi proche d'une gaussienne pour $n = 526$. D'après le théorème central limite avec autonormalisation et la table de la loi normale, on a

$$P\left(\left|\sqrt{n}\frac{\overline{X_n} - p_E}{\sqrt{\overline{X_n}(1-\overline{X_n})}}\right| \leq 1,96\right) \simeq 95\%$$

2. Source : Le devenir des diplômés de Master Professionnel (promotion 2009, enquête 2011). Document OFIP.

qu'on peut réécrire en

$$P\left(p_E \in I\right) \simeq 95\% \quad \text{où} \quad I = \left[\overline{X_n} \pm 1,96 \frac{\sqrt{\overline{X_n}(1-\overline{X_n})}}{\sqrt{n}}\right]$$

Ici $n = 526$ et on observe $\overline{X_n}(\omega) = 487/526$ donc $I(\omega) \simeq [90,34\% \ ; \ 94,83\%]$.

2) Le problème de l'estimation de p_B par intervalle de confiance est analogue à celui de l'estimation de p_E, il suffit juste de remplacer n par 131 et la proportion empirique observée par 110/131. On obtient un intervalle de confiance observé valant $[77,68\% \ ; \ 90,26\%]$. Il est donc assez raisonnable d'affirmer que le taux d'insertion professionnelle est moins élevé en Biologie qu'en Economie. Ceci demanderait à être confirmé en faisant un test.

Ex 5. 1) La loi exponentielle de paramètre 1 a pour espérance 1 et pour variance 1.

2) Puisque les T_i sont i.i.d. de variance strictement positive, le théorème central limite assure que $\frac{\sum_{i=1}^n T_i - nE(T_1)}{\sqrt{n \operatorname{Var}(T_1)}}$ converge en loi vers $\mathcal{N}(0;1)$. En particuler la loi de $\frac{(\sum_{i=1}^{240} T_i) - 240}{\sqrt{240}}$ est proche de $\mathcal{N}(0;1)$, autrement dit la loi de $\sum_{i=1}^{240} T_i$ est proche de $\mathcal{N}(240;240)$.

3) 4 heures et demie, c'est 270 minutes :

$$P\left(\sum_{i=1}^{240} T_i > 270\right) = P\left(\frac{(\sum_{i=1}^{240} T_i) - 240}{\sqrt{240}} > \frac{30}{\sqrt{240}}\right) \simeq P(Z > 1,94) \simeq 1 - 0,9738 = 2,62\%$$

4) Le théorème de Berry-Esséen assure que

$$\left|P\left(\sum_{i=1}^{240} T_i > 270\right) - 0,0262\right| = \left|P\left(\sum_{i=1}^{240} T_i \leq 270\right) - P(Z \leq 1,94)\right| \leq \frac{0,5}{\sqrt{240}}\rho^3$$

Le calcul de la borne nécessite de connaitre le moment centré d'ordre 3 des T_i. On le calcule en utilisant la densité de la loi exponentielle de paramètre 1.

$$\rho^3 = E(|T_1 - 1|^3) = \int_{-\infty}^{+\infty} |x-1|^3 \ f_{T_1}(x) \ dx = \int_0^{+\infty} |x-1|^3 \ e^{-x} \ dx$$

Il faut séparer le cas où $x - 1$ est négatif de celui où il est positif, et un changement de variable en $y = 1 - x$ ou $y = x - 1$ selon le cas simplifie quelque peu le calcul :

$$\rho^3 = \int_0^1 (1-x)^3 \ e^{-x} \ dx + \int_1^{+\infty} (x-1)^3 \ e^{-x} \ dx = \int_0^1 y^3 \ e^{y-1} \ dy + \int_0^{+\infty} y^3 \ e^{-y-1} \ dy$$

Une intégration par partie donne le moment d'ordre 3 de T_1

$$\int_0^{+\infty} x^3 e^{-x} \ dx = [-x^3 e^{-x}]_0^{+\infty} - \int_0^{+\infty} 3x^2(-e^{-x}) \ dx = 3E(T_1^2) = 3(\operatorname{Var}(T_1) + (E(T_1))^2) = 6$$

La même intégration par parties donne

$$\int_0^1 y^3 \ e^y \ dy = [y^3 \ e^y]_0^1 - \int_0^1 3y^2 \ e^y \ dy = e - 3\left([y^2 \ e^y]_0^1 - \int_0^1 2y \ e^y \ dy\right)$$

$$= e - 3\left(e - 2([y \ e^y]_0^1 - \int_0^1 e^y \ dy)\right) = -2e + 6(e - [e^y]_0^1) = 6 - 2e$$

On obtient

$$\rho^3 = \frac{6-2e}{e} + \frac{6}{e} = \frac{12}{e} - 2 \simeq 2,41455$$

ce qui donne une précision de $\frac{0,5}{\sqrt{240}}\rho^3 \simeq 0,0779$. Tout ce qu'on peut affirmer ici, c'est donc que $P\left(\sum_{i=1}^{240} T_i > 270\right)$ est entre 0 et $10,41\%$.

Université Lille1 Sciences et Technologies

U.F.R. de Mathématiques

MASS 3 S5 Estimation **Année 2012-2013 M.Fradon**

Examen, deuxième session, 18 février 2013

Durée : 3 heures.

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

Ex 1. On lance un dé équilibré à 6 faces. On note X_1 le résultat du premier lancer, X_2 le résultat du deuxième lancer, etc. On sait que $E(X_1) = 3,5$. En utilisant l'inégalité de Tchebychev, trouver une valeur n à partir de laquelle la moyenne empirique des résultats a au moins 99% de chances d'être entre 3 et 4 : $P(3 < \overline{X_n} < 4) \geq 99\%$.

Ex 2. Tension de rupture

Le tableau suivant donne les tensions de rupture en kilogramme-force mesurées sur cent segments de fil de nylon. Cette tension de rupture est une variable aléatoire Y de loi inconnue et on peut interpréter ce tableau comme les observations $Y_1(\omega), \ldots, Y_{100}(\omega)$ d'un 100-échantillon de la loi de Y. On s'intéresse à la quantité $\theta = P(Y > 31)$.

TABLE 1 – Un 100-échantillon de tensions de rupture (en kgf) pour un fil de nylon

29,446	31,717	29,818	30,373	31,178	31,146	30,505	28,746	30,037	29,993
28,330	30,734	29,131	29,492	29,554	30,211	29,849	29,147	28,474	31,252
29,003	30,102	28,598	28,552	31,802	30,788	30,060	31,180	30,175	30,556
29,405	31,623	30,786	29,135	29,954	30,782	28,228	30,015	29,856	29,814
30,200	30,688	29,790	30,639	28,846	29,853	29,834	29,140	29,654	29,300
31,195	29,784	31,686	31,001	29,572	29,275	30,160	28,139	29,739	30,869
29,443	27,991	30,020	30,784	28,378	28,554	30,897	28,299	30,723	28,422
30,091	31,060	28,507	31,758	32,052	29,924	30,780	29,086	30,664	30,329
30,922	30,666	29,872	29,035	29,195	31,790	29,412	30,374	28,115	31,782
29,437	31,409	28,381	28,275	29,242	31,132	31,060	29,099	30,127	30,564

1) Des variables aléatoires i.i.d. de loi de Bernoulli de paramètre θ apparaissent naturellement dans ce problème. On les notera $X_1, \ldots, X_n$. Quelle est l'expression des X_i en fonction des Y_i ?

2) On veut une valeur numérique pour estimer θ. Contruire à partir des X_i l'estimateur $\widehat{\theta_n}$ du maximum de vraisemblance. Quelle est ici la valeur observée $\widehat{\theta_n}(\omega)$ de cet estimateur ?

3) $\widehat{\theta_n}$ est-il biaisé ? Converge-t-il si, au lieu de s'en tenir à $n = 100$, on augmente le nombre n de fils testés ? Quel nom porte cette propriété ?

4) Construisez un intervalle de confiance au niveau 95% pour θ en utilisant le théorème central limite (avec ou sans autonormalisation ?). Quelle est la valeur observée de cet intervalle de confiance ?

Ex 3. V suit la loi de Rademacher de paramètre $\frac{1}{2}$:

$$P(V=1) = P(V=-1) = \frac{1}{2}$$

On effectue des tirages indépendants et répétés selon cette loi. On note V_1, V_2, V_3,...les résultats obtenus.

1) La moyenne empirique $\overline{V_n}$ des n premiers tirages converge. Pourquoi ? Vers quelle limite ?

2) Pour $n = 10\,000$ calculer une valeur approximative de $P(|\overline{V_n}| \leq 0,01)$.

3) Quelle est la précision du calcul fait à la question précédente ?

Ex 4. A l'aide d'une calculatrice qui a une fonction "random" (tirage uniforme sur $]0;1[$) on veut simuler la loi de densité $f : \ x \longmapsto \cos(x)\ \mathbf{1}_{]0;\frac{\pi}{2}[}(x)$.

1) Déterminer une densité de probabilité f_W et une constante C telles que $f \leq C\ f_W$ et en déduire une technique pour faire un tirage selon la loi de densité f.

2) Calculer la fonction quantile associée à cette loi et en déduire une autre technique permettant de faire un tirage selon la loi de densité f. Cette technique est-elle meilleure que la précédente ?

Ex 5. La taille des truites

La taille des truites dans une rivière suit une loi gaussienne d'espérance m et de variance σ^2. On mesure la taille de 20 truites péchées dans cette rivière, on note X_1, X_2,...,X_{20} leurs tailles.

1) A partir des X_i, construire un intervalle de confiance pour m au niveau de confiance 95%.

2) Toujours au niveau 95%, construire un intervalle de confiance pour σ.

Données numériques :

Tailles des truites (exprimées en cm) : 20.9 23.0 20.6 23.2 24.6 21.7 21.8 24.2 21.8 21.9 25.1 26.4 23.0 23.7 22.4 17.9 21.9 23.7 23.1 22.0

Somme des tailles : 452.9

Somme des carrés des tailles : 10318.93

Université Lille1 Sciences et Technologies

U.F.R. de Mathématiques

MASS 3 S5 Estimation **Année 2012-2013 M.Fradon**

Corrigé de l'examen de deuxième session, février 2013

Ex 1. Pour utiliser l'inégalité de Tchebychev, on a besoin de la variance des X_i. Comme ils suivent la loi uniforme sur $\{1,2,3,4,5,6\}$:

$$E(X_1^2) = \frac{1}{6}\sum_{k=1}^{6} k^2 = \frac{91}{6} \quad \text{donc } \mathrm{Var}(X_1) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$$

Les X_i ont pour moyenne empirique $\overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i$ qui a pour espérance $\frac{7}{2}$ et pour variance $\frac{35}{12n}$. Tchebychev assure alors que

$$\forall t > 0 \qquad P\left(|\overline{X_n} - \frac{7}{2}| \geq t\right) \leq \frac{35}{12\ n\ t^2}$$

On peut réécrire ceci en

$$\forall t > 0 \qquad P\left(|\overline{X_n} - \frac{7}{2}| < t\right) \geq 1 - \frac{35}{12\ n\ t^2}$$

Or $|\overline{X_n} - \frac{7}{2}| < t \quad \Leftrightarrow \quad \frac{7}{2} - t < \overline{X_n} < \frac{7}{2} + t$ donc en prenant $t = \frac{1}{2}$ on obtient

$$P\left(3 < \overline{X_n} < 4\right) \geq 1 - \frac{35}{12\ n\ (1/2)^2} = 1 - \frac{35}{3\ n}$$

Pour avoir $P(3 < \overline{X_n} < 4) \geq 99\%$ on choisit donc n tel que $\frac{35}{3\ n} \leq 0,01$ i.e. $n \geq 1166,6667$. A partir de 1167 lancers, la moyenne empirique a au moins 99% de chances d'être entre 3 et 4.

Ex 2. 1) Les X_i suivent une loi de Bernoulli de paramètre θ, ce sont donc des indicatrice d'un certain événement de probabilité θ. Puisque $\theta = P(Y_i > 31)$ pour chaque Y_i, le choix de $X_i = \mathbf{1}_{Y_i > 31}$ donne bien des v.a. indépendantes (les Y_i le sont) et de même loi $\mathcal{B}er(\theta)$.

2) Pour estimer θ par maximum de vraisemblance, on calcule d'abord la vraisemblance

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

et on passe au logarithme

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

On dérive

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

et on trouve que la vraisemblance est maximale en $\theta = \frac{1}{n}\sum_{i=1}^n x_i$. Donc $\widehat{\theta_n} = \overline{X_n}$.
Utilisons les observations fournies : $X_1(\omega) = 0$, $X_2(\omega) = 1$, $X_3(\omega) = 0$, $X_4(\omega) = 0$, $X_5(\omega) = 1$, $X_6(\omega) = 1$,...On constate que $\widehat{\theta_n}(\omega) = \overline{X_n}(\omega) = 18/100$.

3) $E(\overline{X_n}) = E(X_1) = \theta$. $\widehat{\theta_n}$ est donc un estimateur sans biais de θ. De plus, la loi forte des grands nombres assure que $\overline{X_n}$ converge presque surement vers $E(X_1)$, autrement dit que $\widehat{\theta_n}$ converge presque surement vers θ. C'est donc un estimateur fortement consistant de θ.

4) On a $n = 100$ données, on va utiliser le théorème central limite. La variance $\theta(1-\theta)$ des X_i étant inconnue, on va se servir de l'autonormalisation :

$$P\left(\left|\sqrt{n}\frac{\overline{X_n} - \theta}{\sqrt{\overline{X_n}(1-\overline{X_n})}}\right| \leq 1,96\right) \simeq 95\%$$

c'est-à-dire $P(\theta \in I) \simeq 95\%$ où $I = \left[\overline{X_n} \pm 1,96\frac{\sqrt{\overline{X_n}(1-\overline{X_n})}}{\sqrt{n}}\right]$.
On observe $\overline{X_n}(\omega) = 0,18$ donc $I(\omega) \simeq [0,104\ ;\ 0,255]$.

Ex 3. 1) La loi de Rademacher est intégrable : $E(V) = \frac{1}{2} - \frac{1}{2} = 0$. D'après la loi forte des grands nombres, la moyenne empirique $\overline{V_n}$ converge presque sûrement vers l'espérance de V, donc vers zéro.

2) On va utiliser le théorème central limite pour calculer $P(|\overline{V_n}| \leq 0,01)$. Pour cela, on aura besoin de la variance de la loi de Rademacher, qui est facile à calculer puisque son support est réduit à deux valeurs opposées : $P(V^2 = 1) = 1$ donc $E(V^2) = 1$ et donc $\mathrm{Var}(V) = E(V^2) - 0^2 = 1$. Par conséquent, le théorème central limite donne pour $Z \sim \mathcal{N}(0;1)$:

$$P\left(\left|\overline{V_n}\right| \leq 0,01\right) = P\left(\left|\sqrt{n}\frac{\overline{V_n} - 0}{\sqrt{1}}\right| \leq 0,01\sqrt{n}\right) \simeq P\left(|Z| \leq 0,01\sqrt{n}\right)$$

$n = 10\,000$ i.e.$0,01\sqrt{n} = 1$ et $P(|Z| \leq 1) \simeq 2 \times 0,8414 - 1 \simeq 68,3\%$.

3) La précision de ce calcul approximatif est donnée par le théorème de Berry-Esséen, et pour l'utiliser on a besoin du moment centré d'ordre 3 de V :

$$\rho^3 = E(|V - E(V)|^3) = E(|V|^3) = 1$$

D'après le théorème de Berry-Esséen :

$$\left|P\left(\left|\overline{V_n}\right| \leq 0,01\right) - P(|Z| \leq 1)\right| \leq 2\frac{0,5}{\sqrt{10\,000}}\frac{\rho^3}{\mathrm{Var}(V)^{3/2}} = 0,01$$

On conclut que $P\left(\left|\overline{V_n}\right| \leq 0,01\right)$ est entre $67,3\%$ et $69,3\%$

Ex 4. 1) La densité $f :\ x \longmapsto \cos(x)\ \mathbf{1}_{]0;\frac{\pi}{2}[}(x)$ est majorée par 1 sur $]0;\frac{\pi}{2}[$ et nulle en dehors, donc

$$\forall x \in \mathbb{R} \quad f(x) \leq \mathbf{1}_{]0;\frac{\pi}{2}[}(x) = C\ f_W(x) \quad \text{où} \quad f_W = \frac{2}{\pi}\mathbf{1}_{]0;\frac{\pi}{2}[} \text{ et } C = \frac{\pi}{2}$$

On utilise la méthode du rejet. On tire d'abord un nombre W_1 au hasard (uniformément) entre 0 et $\frac{\pi}{2}$, puis on tire un U_1 au hasard entre 0 et 1 (uniformément, et indépendamment de W_1). La condition d'acceptation est $C\ f_W(W_1)\ U_1 \leq f(W_1)$ i.e.

$$\frac{\pi}{2}\ \frac{2}{\pi}\mathbf{1}_{]0;\frac{\pi}{2}[}(W_1)\ U_1 \leq \cos(W_1)\ \mathbf{1}_{]0;\frac{\pi}{2}[}(W_1) \quad \text{i.e.} \quad U_1 \leq \cos(W_1)$$

Si elle est satisfaite on accepte W_1 sinon on le rejette et on recommence la procédure avec un W_2 et un U_2, éventuellement un W_3 et un U_3, etc jusqu'à ce qu'on accepte un W_i. La loi de cette valeur acceptée est bien celle de densité f.

2) La fonction de répartition F associée à la loi de densité f est

$$\forall t \in \mathbb{R} \quad F(t) = \int_{-\infty}^{t} \cos(x)\ \mathbf{1}_{]0;\frac{\pi}{2}[}(x)\ dx = \sin(t)\mathbf{1}_{]0;\frac{\pi}{2}[}(t) + \mathbf{1}_{]\frac{\pi}{2};+\infty[}(t)$$

donc la fonction quantile associée à cette loi est

$$\forall u \in]0;1[\quad F^{-1}(u) = \inf\{t \in \mathbb{R}\ ;\ F(t) \geq u\} = \inf\{t > 0\ ;\ \sin(t) \geq u\} = \arcsin(u)$$

Si U suit la loi uniforme sur $]0;1[$, $\arcsin(U)$ aura donc pour densité f. Cette technique a l'avantage de simuler la loi de densité f en n'utilisant qu'un seul tirage de loi uniforme (la fonction arcsin est préprogrammée dans la plupart des langages).

Université Lille1
Sciences et Technologies

U.F.R. de Mathématiques

MASS 3 S5 Estimation **Année 2012-2013 M.Fradon**

Partiel, 15 novembre 2012

Durée : 2 heures.

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

Ex 1. 1) En France, 37% des gens sont du groupe sanguin O^+. Quelle est la probabilité que parmi 300 donneurs de sang, il y en ait au plus 105 du groupe O^+ ? On calculera une valeur approchée de cette probabilité, puis on donnera la précision de la valeur calculée.

2) Seulement 2% des gens sont du groupe sanguin B^-. Quelle est approximativement la probabilité que parmi 300 donneurs de sang, il y en ait au plus 3 du groupe B^- ?

Ex 2. Des petits crustacées vivent sur une côte. Certains sont porteurs de parasites. La proportion p de crustacés parasités est inconnue.

1) A marée basse, un pêcheur ramasse au hasard quelques petits crustacés. Son seau contient 28 crustacés, dont X sont parasités. Construire à partir de X un intervalle de confiance I au niveau 90% pour l'estimation de la proportion p de crustacés parasités.

2) On compte les crustacés parasités dans le seau du pêcheur : on en trouve 10. Quelle valeur a pris l'intervalle de confiance I ? Pourquoi cette estimation de p n'est-elle pas satisfaisante ?

3) On organise une étude scientifique du parasitage des crustacés. 3600 crustacés vont être capturés, observés, puis relâchés. On note Y le nombre de crustacés parasités qu'on trouvera parmi eux. Construire à partir de Y un intervalle de confiance J pour l'estimation de p (toujours au niveau 90%).

4) Sur les 3600 crustacés, 1176 portaient des parasites. Quelle valeur a pris J ?

Ex 3. La variable aléatoire X a pour fonction de répartition F définie par

$$F(t) = \frac{(t+1)^2}{4}\mathbf{1}_{[0;1[}(t) + \mathbf{1}_{[1;+\infty[}(t)$$

On tire 10 000 variables aléatoires X_i indépendantes de même loi que X. Le but est de calculer la probabilité que leur somme soit entre 4094 et 4207.

1) Tracer le graphe de F.

2) Déterminer la fonction quantile de X. En déduire un moyen de simuler X.

3) Simuler X par la méthode du rejet décrite en cours n'est pas possible. Pourquoi ?

4) Prouver que X a pour espérance $\frac{5}{12}$.

5) Déterminer la fonction de répartition de la variable aléatoire X^2. En déduire que X a pour variance $\frac{17}{144}$.

6) Calculer la probabilité que $4094 \leq \sum_{i=1}^{10\,000} X_i \leq 4207$.

Université Lille1 Sciences et Technologies

U.F.R. de Mathématiques

MASS 3 S5 Estimation **Année 2012-2013 M.Fradon**

Corrigé du partiel du 15 novembre 2012

Ex 1. 1) Il est raisonnable de supposer l'indépendance entre les donneurs de sang. En notant X_i l'indicatrice du fait que la personne numéro i est du groupe O^+, les X_i sont alors indépendantes toutes de loi $\mathcal{B}er(0,37)$. Leur variance est $0,37 \times 0,63 = 0,2331 > 0$. Le théorème central limite assure que leur moyenne empirique satisfait

$$\forall t \in \mathbb{R} \quad \lim_{n\to+\infty} P(\sqrt{n}\frac{\overline{X}_n - 0,37}{\sqrt{0,2331}} \leq t) = P(Z \leq t)$$

où Z désigne une variable aléatoire de loi gaussienne centrée réduite.

Le nombre de gens du groupe O^+ parmi les donneurs de sang est $\sum_{i=1}^n X_i = n\overline{X}_n$. Comme $n = 300$ est relativement grand

$$P(300\overline{X}_{300} \leq 105) = P(\sqrt{300}\frac{\overline{X}_{300} - 0,37}{\sqrt{0,2331}} \leq \sqrt{300}\frac{\frac{105}{300} - 0,37}{\sqrt{0,2331}}) \simeq P(Z \leq -0,72) \simeq 1-0,7642 \simeq 23,6\%$$

La précision de l'approximation faite en utilisant le théorème central limite peut être évaluée grâce au théorème de Berry-Esséen, qui est bien utilisable ici puisqu'on part d'une suite de Bernoulli indépendantes de paramètre différent de 0 et de 1.

$$\left|P(300\overline{X}_{300} \leq 105) - P(Z \leq -0,72)\right| \leq \frac{0,5}{\sqrt{300}}\ \frac{0,37^2 + 0,63^2}{\sqrt{0,37 \times 0,63}} \simeq 0,032$$

On conclut que la probabilité qu'il y ait au plus 105 donneurs du groupe O^+ est entre 20,4% et 26,8%.

2) On suppose toujours l'indépendance entre les donneurs de sang, et on note Y_i l'indicatrice du fait que la personne numéro i est du groupe B^-. Comme les Y_i sont i.i.d. de loi $\mathcal{B}er(0,02)$, leur somme suit la binomiale $\mathcal{B}in(300;0,02)$. Puisque $300 \geq 100$, on approche $\mathcal{B}in(300;0,02)$ par la loi de Poisson de paramètre $300 \times 0,02 = 6$ qui est inférieur à 10.

$$P(\sum_{i=1}^{300} Y_i \leq 3) \simeq e^{-6}\sum_{k=0}^{3}\frac{6^k}{k!} = e^{-6}(1+6+18+36) \simeq 15,1\%$$

La binomiale $\mathcal{B}in(300;0,02)$ qui est mal approximée par une gaussienne : la borne de Berry-Esséen vaut $\frac{0,5}{\sqrt{300}}\ \frac{0,02^2+0,98^2}{\sqrt{0,02\times0,98}} \simeq 0,20$ ici !

Remarque : la valeur exacte est $14,851\%$. *Attention, l'approximation gaussienne via le théorème central limite donnait* $10,75\%$. *La* $\mathcal{B}in(300;0,02)$ *est mal approximée par une gaussienne : la borne de Berry-Esséen vaut* $\frac{0,5}{\sqrt{300}}\ \frac{0,02^2+0,98^2}{\sqrt{0,02\times0,98}} \simeq 0,20$ *ici ! Comme on l'a vu, il vaut mieux utiliser l'approximation poissonienne des binomiales quand leur deuxième paramètre est très grand ou très petit.*

Ex 2. 1) En supposant l'indépendance entre les crustacés, le nombre X de parasités parmi les 28 suit la loi $\mathcal{B}in(28;p)$. D'après l'inégalité de Tchebychev :
$P(|X-28p|\geq t)\leq \frac{28p(1-p)}{t^2}$ pour chaque t strictement positif.
En majorant $p(1-p)$ par $\frac{1}{4}$, on obtient $P(|X-28p|\geq t)\leq \frac{28}{4t^2}$.
On choisit t pour avoir $\frac{28}{4t^2}=0,1$: $\quad t=\sqrt{\frac{28}{4\times 0,1}}=\sqrt{70}$.

$$P(|\frac{X}{28}-p|\geq \frac{\sqrt{70}}{28})\leq 0,1 \quad \text{i.e.} \quad P(|\frac{X}{28}-p|<\frac{\sqrt{70}}{28})\geq 0,9$$

donc $I=]\frac{X}{28}-\frac{\sqrt{70}}{28};\frac{X}{28}+\frac{\sqrt{70}}{28}[\simeq]\frac{X}{28}-0,3;\frac{X}{28}+0,3[$ est un intervalle de confiance au niveau 90% pour l'estimation de la proportion p de crustacés parasités.

2) Ce jour-là, on a observé $X(\omega)=10$ donc $I(\omega)=]5,8\%;65,6\%[$. Cet intervalle est tellement large qu'il est presque inutilisable.

3) Les observations vont maintenant porter sur un nombre beaucoup plus grands de crustacés, on va pouvoir utiliser le théorème central limite. Le nombre Y de crustacés parasités suit la loi $\mathcal{B}in(3600;p)$. La variance, qui dépend de p, est inconnue. Il faut donc utiliser le théorème central limite avec autonormalisation. On note Z une variable aléatoire de loi gaussienne centrée réduite et on lit dans la table de valeurs numérique que $P(Z\leq 1,645)\simeq 95\%$.

$$P\left(\left|\sqrt{3600}\frac{\frac{Y}{3600}-p}{\sqrt{\frac{Y}{3600}(1-\frac{Y}{3600})}}\right|\leq 1,645\right)\simeq P(|Z|\leq 1,645)\simeq 90\%$$

et par conséquent $P\left(\left|\frac{Y}{3600}-p\right|\leq \frac{1,645}{60}\sqrt{\frac{Y}{3600}(1-\frac{Y}{3600})}\right)\simeq 90\%$

L'intervalle de confiance cherché est donc

$$J=\left[\frac{Y}{3600}-\frac{1,645}{60}\sqrt{\frac{Y}{3600}(1-\frac{Y}{3600})};\frac{Y}{3600}+\frac{1,645}{60}\sqrt{\frac{Y}{3600}(1-\frac{Y}{3600})}\right]$$

4) Sur les 3600 crustacés, $Y(\omega)=1176$ portaient des parasites. On observe pour ce tirage $J(\omega)\simeq[31,3\%;34,0\%]$

Ex 3. 1)

2) Pour tout u de $]0;1[$, $F^{-1}(u)=\inf\{t\in\mathbb{R};F(t)\geq u\}$.
F est nulle sur $]-\infty;0[$ et supérieure ou égale à $\frac{1}{4}$ sur $[0;+\infty[$ donc pour u dans $]0;\frac{1}{4}[$ on a $\{t\in\mathbb{R};F(t)\geq u\}=[0;+\infty[$ et $F^{-1}(u)=\inf[0;+\infty[=0$.
Pour u dans $[\frac{1}{4};1[$ on a

$F^{-1}(u) = \inf\{t \in [0;1[; \frac{(t+1)^2}{4} \geq u\} = \inf\{t \in [0;1[; t+1 \geq \sqrt{4u}\} = \sqrt{4u} - 1$.
Donc on peut simuler X par

$$F^{-1}(U) = (2\sqrt{U} - 1)\mathbf{1}_{U > \frac{1}{4}} \quad \text{où} \quad U \sim \mathcal{U}nif(]0;1[)$$

3) La méthode du rejet décrite en cours nous donne un moyen de simuler les variables aléatoires réelles discrètes ou à densité et les vecteurs aléatoires à densité. X est une variable aléatoire réelle qui n'est pas à densité (car $P(X = 0) > 0$) et qui n'est pas discrète non plus (puisque $\sum_{x \in \mathbb{R}} P(X = x) = P(X = 0) < 1$).

4) $P(X < 0) = \lim_{t \to 0, t<0} F(t) = 0$ donc

$$E(X) = \int_0^{+\infty} 1 - F(t)\ dt = \int_0^1 1 - \frac{(t+1)^2}{4}\ dt + \int_1^{+\infty} 1 - 1\ dt = 1 - \frac{1}{4}\int_0^1 t^2 + 2t + 1\ dt$$

$$E(X) = 1 - \frac{1}{4}\left[\frac{t^3}{3} + t^2 + t\right]_0^1 = 1 - \frac{1}{4}(\frac{1}{3} + 2) = \frac{5}{12}$$

5) Si t est strictement négatif $P(X^2 \leq t) = 0$.
$P(X^2 \leq 0) = P(X = 0) = \frac{1}{4}$.
Si t est strictement positif inférieur à 1, on trouve $P(X^2 \leq t) = P(X \leq \sqrt{t}) = \frac{(\sqrt{t}+1)^2}{4}$.
Et pour $t \geq 1$ on aura aussi $P(X^2 \leq t) = P(X \leq \sqrt{t}) = 1$ puisque $\sqrt{t} \leq 1$.
Donc X^2 a pour fonction de répartition $F_{X^2} : t \longmapsto \frac{(\sqrt{t}+1)^2}{4}\mathbf{1}_{[0;1[}(t) + \mathbf{1}_{[1;+\infty[}(t)$.
On en déduit son espérance : $E(X^2) = \int_0^{+\infty} 1 - F_{X^2}(t)\ dt = \int_0^1 1 - \frac{(\sqrt{t}+1)^2}{4}\ dt$

$$E(X^2) = 1 - \frac{1}{4}\int_0^1 t + 2\sqrt{t} + 1\ dt = 1 - \frac{1}{4}\left[\frac{t^2}{2} + 2\frac{2}{3}t^{3/2} + t\right]_0^1 = 1 - \frac{1}{4}(\frac{1}{2} + \frac{4}{3} + 1) = 1 - \frac{1}{4} \times \frac{17}{6} = \frac{7}{24}$$

ce qui donne $Var(X) = E(X^2) - (E(X))^2 = \frac{7}{24} - \frac{25}{144} = \frac{17}{144}$.

6) Les X_i sont i.i.d. de variance non-nulle. D'après le théorème central limite

$$\sqrt{n}\frac{\overline{X}_n - \frac{5}{12}}{\sqrt{\frac{17}{144}}} \xrightarrow[n \to +\infty]{Loi} Z \sim \mathcal{N}(0,1)$$

et donc pour $n = 10\,000$: $P\left(4094 \leq \sum_{i=1}^{n} X_i \leq 4207\right) = P\left(0,4094 \leq \overline{X}_n \leq 0,4207\right)$

$$= P\left(\sqrt{n}\frac{0,4094 - \frac{5}{12}}{\sqrt{\frac{17}{144}}} \leq \sqrt{n}\frac{\overline{X}_n - \frac{5}{12}}{\sqrt{\frac{17}{144}}} \leq \sqrt{n}\frac{0,4207 - \frac{5}{12}}{\sqrt{\frac{17}{144}}}\right) \simeq P\left(100\frac{0,4094 - \frac{5}{12}}{\sqrt{\frac{17}{144}}} \leq Z \leq 100\frac{0,4207 - \frac{5}{12}}{\sqrt{\frac{17}{144}}}\right)$$

La probabilité cherchée est égale à $P(-2,11 \leq Z \leq 1,17) \simeq 0,8790 - (1 - 0,9826) = 86,2\%$.