Université Lille1
Sciences et Technologies

U.F.R. de Mathématiques

MASS 3 S5 Estimation **Année 2014-2015 M.Fradon**

Examen, 18 décembre 2014

Durée : 3 heures.

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

Ex 1.

Une route départementale très fréquentée comporte une ligne droite longue de 7 km. On modélise la durée de parcours sur ce tronçon par une loi gaussienne d'espérance m et d'écart-type σ. On veut estimer les paramètres inconnus m et σ en utilisant les temps de parcours T_1, T_2,..., T_{30} de trente voitures choisies au hasard.

1) Construire à partir de T_1, T_2,..., T_{30} des intervalle de confiance pour l'estimation de m et σ, au niveau de confiance 90%.

2) On mesure les temps de parcours en minutes de trente voitures prises au hasard. On obtient les observations suivantes : 3.90 4.76 3.51 4.20 3.51 4.27 4.24 3.79 3.69 4.44 4.44 3.78 4.05 4.66 5.23 4.00 4.02 4.42 3.94 4.05 4.20 4.29 3.63 4.89 4.90 4.95 4.27 3.79 4.57 3.83 (la somme de ces valeurs est 126.22 et la somme de leurs carrés est 536.92).

Calculer les valeurs observées des intervalles de confiance construits. Peut-on raisonnablement penser que la vitesse limite de 90 km/h est respectée dans cette ligne droite ?

Ex 2. Saisie informatique

Un opérateur lit des valeurs sur des fiches et les tape au clavier. En cas d'erreur de lecture ou de faute de frappe, la donnée tapée est fausse. Chaque donnée, indépendamment des autres, a une probabilité $0,5\%$ d'être fausse. L'opérateur tape 2400 données au total.

1) Quelle est la loi du nombre N de données fausses ? Calculer son espérance et sa variance.

2) On veut connaitre la probabilité qu'il n'y ait aucune erreur. Donner une expression de $P(N = 0)$ et calculer sa valeur numérique.

3) En utilisant le théorème central limite pour les binomiales, calculer approximativement la probabilité qu'au moins 2380 données sur les 2400 soient correctes.

4) Quelle est la précision de l'approximation faite à la question précédente ? En déduire un encadrement de la probabilité qu'au moins 2380 données soient correctes.

5) En utilisant l'inégalité de Tchebychev, minorer la probabilité qu'au moins 2380 données soient correctes. L'inégalité de Tchebychev donne-t-elle de meilleurs résultats que l'approximation gaussienne ici ?

Ex 3.

On définit la fonction $f \ : x \longmapsto |x| \ \mathbf{1}_{-1 \leq x \leq 1}$ sur $\mathbb{R}$.

1) Prouver que f est une densité de probabilité. On note μ la loi de densité f.

2) Expliquer ce qu'on doit faire pour simuler la loi μ par la méthode du rejet (quels tirages ? quelle condition d'arrêt ?).

3) Expliquer ce qu'on doit faire pour simuler la loi μ par inversion de la fonction de répartition (indiquer la fonction quantile et le tirage à faire).

Ex 4.

La variable aléatoire X a pour densité

$$\forall x \in \mathbb{R} \quad f_\alpha(x) = (\alpha + 1) \ x^\alpha \ \mathbf{1}_{0 \leq x \leq 1}$$

Le paramètre $\alpha > 0$ est inconnu. On veut l'estimer à partir des résultats X_1, X_2,..., X_{625} de 625 tirages indépendants de même loi que X.

1) Calculer l'espérance de X en fonction de α. On la note m_α.

2) Construire à partir de X_1, X_2,..., X_{625} un intervalle de confiance pour l'estimation de m_α au niveau de confiance 96%.

3) En déduire un intervalle de confiance pour l'estimation de α, toujours au niveau de confiance 96%.

4) A partir de n tirages X_1, X_2,..., X_n i.i.d. de même loi que X, construire l'estimateur du maximum de vraisemblance pour α. On le note $\hat{\alpha}_n$.

5) Prouver que $\hat{\alpha}_n$ est un estimateur fortement consistant de α.

6) *Question bonus, hors barème*

On note $Y_i = \ln(X_i)$ pour i de 1 à 625. Calculer la variance des Y_i et, en utilisant le théorème central limite, construire un autre intervalle de confiance à 96% pour l'estimation de α.

Université Lille1 Sciences et Technologies

U.F.R. de Mathématiques

MASS 3 S5 Estimation **Année 2014-2015 M.Fradon**

Corrigé de l'examen du 16 décembre 2013

Ex 1. 1) On peut supposer les 30 durées de parcours $T_1,\dots,T_{30}$ indépendantes. Elles suivent toutes la même loi gaussienne. L'écart-type est non-nul car les durées de parcours observées ne sont pas toutes égales. En notant $\overline{T_{30}}$ et V_{30} l'espérance empirique et la variance empirique, le théorème de Student permet d'affirmer que

$$\sqrt{29}\,\frac{\overline{T_{30}}-\mu}{\sqrt{V_{30}}} \sim \text{Student}(29) \quad \text{et} \quad \frac{30\,V_{30}}{\sigma^2} \sim \chi^2(29)$$

La table de la Student(29) et la symétrie de cette loi donnent :

$$P\left(\left|\sqrt{29}\,\frac{\overline{T_{30}}-\mu}{\sqrt{V_{30}}}\right| \leq 1,699\right) \simeq 90\%$$

On obtient l'intervalle de confiance sur μ : $I = \left[\overline{T_{30}} \pm 1,699\frac{\sqrt{V_{30}}}{\sqrt{29}}\right]$

On utilise ensuite la table du $\chi^2(29)$ qui donne :

$$P\left(17,708 \leq \frac{30\,V_{30}}{\sigma^2} \leq 42,557\right) = P\left(\frac{30\,V_{30}}{17,708} \geq \sigma^2 \geq \frac{30\,V_{30}}{42,557}\right) \simeq 90\%$$

d'où l'intervalle de confiance sur la volatilité σ : $J = \left[\sqrt{\frac{30\,V_{30}}{42,557}}\ ;\ \sqrt{\frac{30\,V_{30}}{17,708}}\right]$.

2) $\overline{T_{30}}(\omega) \simeq 4,207$ et $V_{30}(\omega) = 536,92/30 - (\overline{T_{30}}(\omega))^2 \simeq 0,197$ donc

$$I(\omega) \simeq [4,06\ ;\ 4,35] \quad \text{et} \quad J(\omega) \simeq [0,372\ ;\ 0,578]$$

Un temps de parcours de $4,06$ minutes pour 7 km correspond à une vitesse de $7/(4,06/60) \simeq 103,45$ km/h. Et un temps de parcours de $4,35$ minutes correspond à une vitesse de $7/(4,35/60) \simeq 96,55$ km/h. L'intervalle de confiance observé permet donc raisonnablement d'affirmer que la vitesse moyenne sur ce tronçon est entre 96 et 104 km/h : la limite de 90 km/h n'est pas respectée.

Ex 2. Saisie informatique

1) Chacune des 2400 données, indépendamment des autres, a une probabilité $0,5\%$ d'être fausse. Donc le nombre N de données fausses suit la loi binomiale de paramètres 2400 et $0,005$. Son espérance est $2400 \times 0,005 = 12$ et sa variance est $2400 \times 0,005 \times 0,995 = 11,94$.

2) La probabilité exacte qu'il n'y ait aucune erreur est $P(N = 0) = 0,995^{2400} \simeq 5,962.10^{-6}$.

3) La probabilité qu'au moins 2380 données sur les 2400 soient correctes vaut

$$P(N \leq 20) = P\left(\frac{N-12}{\sqrt{11,94}} \leq \frac{20-12}{\sqrt{11,94}}\right)$$

D'après le théorème central limite, ceci est proche de

$$P(N \leq 20) = P\left(\frac{N-12}{\sqrt{11,94}} \leq 2,315\right) \simeq 98,965\%$$

4) Le théorème de Berry-Esséen assure que

$$|P(N \leq 20) - 0,98965| \leq \frac{(0,005)^2 + (0,995)^2}{2\sqrt{2400 \times 0,005 \times 0,995}} \simeq 0,14326$$

$P(N \leq 20)$ est donc entre $0,98965 - 0,14326 = 84,64\%$ et 100%.

5) On utilise l'inégalité de Tchebychev : $P(N \geq 21) = P(N - 12 \geq 9)$ donc

$$P(N \geq 21) \leq P\left(|N-12| \geq 9\right) \leq \frac{11,94}{9^2} \simeq 14,74\%$$

On en déduit que $P(N \leq 20) = 1 - P(N \geq 21) \geq 85,36\%$.

L'inégalité de Tchebychev donne donc un résultat un peu plus précis que l'approximation gaussienne ici, à condition de l'utiliser au mieux en majorant $P(N \geq 21)$. Si on se contente de majorer $P(N \geq 20)$ on obtient

$$P(N \geq 20) = P(N - 12 \geq 8) \leq P\left(|N-12| \geq 8\right) \leq \frac{11,94}{8^2} \simeq 18,66\%$$

On en déduit que $P(N \leq 20) \geq 1 - P(N \geq 20) \geq 84,34\%$. Ce résultat est un peu moins précis que celui obtenu grâce à l'approximation gaussienne.

Ex 3.

1) La fonction $f : x \longmapsto |x| \ \mathbf{1}_{-1 \leq x \leq 1}$ est positive sur $\mathbb{R}$.

$$\int_{\mathbb{R}} f(x)\,dx = \int_{\mathbb{R}} |x| \ \mathbf{1}_{-1 \leq x \leq 1}\,dx = \int_{-1}^{1} |x|\,dx = \int_{-1}^{0} -x\,dx + \int_{0}^{1} x\,dx = -\left[\frac{x^2}{2}\right]_{-1}^{0} + \left[\frac{x^2}{2}\right]_{0}^{1} = \frac{1}{2} + \frac{1}{2} = 1$$

donc f est une densité de probabilité.

2) On sait tirer W selon la loi $\mathcal{U}nif([-1;1])$, par exemple en tirant $W = 2V - 1$ où $V \sim \mathcal{U}nif([0;1])$. La v.a. W a pour densité $f_W = \frac{1}{2}\mathbf{1}_{[-1;1]}$ et les valeurs prises par f sur $[-1;1]$ sont entre 0 et 1 donc $f \leq 2f_W$, sur tout $\mathbb{R}$ puisque les deux densités sont nulles en dehors de $[-1;1]$. On peut donc utiliser la méthode du rejet :

— On tire W_1 de même loi que W et $U_1 \sim \mathcal{U}nif([0;1])$ indépendante de W_1.
— On regarde si $2f_W(W_1)U_1 \leq f(W_1)$, i.e. si $U_1 \leq |W_1|$.
— Si cette inégalité n'est pas vérifiée, on "rejette" W_1, c'est-à-dire qu'on tire, indépendamment des v.a. précédentes, W_2 de même loi que W et $U_2 \sim \mathcal{U}nif([0;1])$ indépendante de W_2.
— On regarde si $U_2 \leq |W_2|$. Si oui on accepte W_2, si non on le rejette, etc

La valeur acceptée (le dernier W_i tiré) suit la loi de densité f.

3) La fonction de répartition F de μ est une primitive de la densité f :

$$\forall t \in \mathbb{R} \quad F(t) = \int_{-\infty}^{t} |x| \ \mathbf{1}_{-1\leq x\leq 1} \ dx$$

— Si $t \leq -1$ on obtient $F(t) = 0$.
— Si $-1 \leq t \leq 0$ alors $F(t) = \int_{-1}^{t} -x \ dx = -\left[\frac{x^2}{2}\right]_{-1}^{t} = \frac{1}{2} - \frac{t^2}{2}$.
— Si $0 \leq t \leq 1$ on a $F(t) = \int_{-1}^{0} -x \ dx + \int_{0}^{t} -x \ dx = \frac{1}{2} + \left[\frac{x^2}{2}\right]_{0}^{t} = \frac{1}{2} + \frac{t^2}{2}$.
— Si $t \geq 1$ alors $F(t) = 1$.

On calcule alors la fonction quantile : $F^{-1}(u) = \inf\{t \in \mathbb{R} \ ; \ F_(t) \geq u\}$ pour $u \in]0;1[$.
— Si $u \leq \frac{1}{2}$

$$F^{-1}(u) = \inf\{t \in [-1;0] \ ; \ \frac{1}{2} - \frac{t^2}{2} \geq u\} = \inf\{t \in [-1;0] \ ; \ 1 - 2u \geq t^2\} = -\sqrt{1-2u}$$

— Si $u \geq \frac{1}{2}$

$$F^{-1}(u) = \inf\{t \in [0;1] \ ; \ \frac{1}{2} + \frac{t^2}{2} \geq u\} = \inf\{t \in [0;1] \ ; \ t^2 \geq 2u - 1\} = \sqrt{2u-1}$$

On effectue un tirage de loi μ en tirant $-\mathbf{1}_{U\leq 1/2}\sqrt{1-2U} + \mathbf{1}_{U>1/2}\sqrt{2U-1}$ où $U \sim \mathcal{U}nif([0;1])$.

Ex 4.

1) f_α est nulle en dehors de $[0;1]$ donc $P(X \in [0;1]) = 1$. Par conséquent, X a des moments de tous ordres.

$$E(X) = \int_{\mathbb{R}} x \ f_\alpha(x) \ dx = (\alpha+1)\int_0^1 x^{\alpha+1} \ dx = (\alpha+1)\left[\frac{x^{\alpha+2}}{\alpha+2}\right]_0^1 = \frac{\alpha+1}{\alpha+2} = 1 - \frac{1}{\alpha+2} =: m_\alpha$$

2) Les X_i sont des v.a. i.i.d. Leur variance est strictement positive puisqu'elles sont à densité. D'après le théorème central limite avec autonormalisation, en notant V_{625} leur variance empirique, on sait que la loi de $\sqrt{625}\frac{\overline{X_{625}} - m_\alpha}{\sqrt{V_{625}}}$ est proche de celle de la gaussienne centrée réduite Z et donc

$$P\left(\left|\sqrt{625}\frac{\overline{X_{625}} - m_\alpha}{\sqrt{V_{625}}}\right| \leq 2,05\right) \simeq P\left(|Z| \leq 2,05\right) \simeq 96\%$$

ce qui donne

$$P\left(m_\alpha \in I\right) \simeq 96\% \quad \text{pour} \quad I = \left[\overline{X_{625}} \pm 2,05\frac{\sqrt{V_{625}}}{\sqrt{625}}\right]$$

3)

$$\overline{X_{625}} - 2,05\frac{\sqrt{V_{625}}}{\sqrt{625}} \leq m_\alpha = 1 - \frac{1}{\alpha+2} \leq \overline{X_{625}} + 2,05\frac{\sqrt{V_{625}}}{\sqrt{625}}$$

est équivalent à

$$1 - \overline{X_{625}} + 2,05\frac{\sqrt{V_{625}}}{25} \geq \frac{1}{\alpha+2} \geq 1 - \overline{X_{625}} - 2,05\frac{\sqrt{V_{625}}}{25}$$

et aussi à

$$\left(1 - \overline{X_{625}} + 2,05\frac{\sqrt{V_{625}}}{25}\right)^{-1} - 2 \leq \alpha \leq \left(1 - \overline{X_{625}} - 2,05\frac{\sqrt{V_{625}}}{25}\right)^{-1} - 2$$

Cette inégalité est donc réalisée avec probabilité 96% ce qui fournit un intervalle de confiance pour l'estimation de α.

4) La vraisemblance, fonction des observations x_1, x_2,..., x_n et du paramètre inconnu α, vaut

$$L(x_1,\dots,x_n,\alpha)=\prod_{i=1}^n f_\alpha(x_i)=\prod_{i=1}^n\left((\alpha+1)\ x_i^\alpha\ \mathbf{1}_{0\le x_i\le 1}\right)=(\alpha+1)^n(\mathbf{1}_{\forall i\le n\quad x_i\in[0;1]})\left(\prod_{i=1}^n x_i\right)^\alpha$$

Puisque $P(X\in[0;1])=1$, on a une probabilité 1 que toutes les observations soient dans $[0;1]$. On peut donc supposer les x_i tous dans $[0;1]$.

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

La dérivée par rapport à α vaut

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

Elle est positive ssi

$$\frac{1}{\alpha+1}\ge\frac{-1}{n}\sum_{i=1}^n\ln(x_i)\quad\text{ i.e. }\quad\alpha+1\le\frac{-n}{\sum_{i=1}^n\ln(x_i)}$$

donc la vraisemblance atteint son unique maximum en $\alpha=\frac{-n}{\sum_{i=1}^n\ln(x_i)}-1$. L'estimateur du maximum de vraisemblance du paramètre inconnu α est donc $\widehat{a_n}=\frac{-n}{\sum_{i=1}^n\ln(X_i)}-1$.

5) Les X_i sont i.i.d. donc les $Y_i=\ln(X_i)$ le sont aussi. La fonction $x\longmapsto\ln(x)\ x^\alpha$ admet un prolongement continu à $[0;1]$ puisque $\lim_{x\to0^+}\ln(x)\ x^\alpha=0$. Par conséquent, la v.a. $\ln(X)$ est intégrable.

$$E(\ln(X))=\int_{\mathbb{R}}\ln(x)\ (\alpha+1)\ x^\alpha\ \mathbf{1}_{0\le x\le1}\ dx=\int_0^1\ln(x)\ (\alpha+1)\ x^\alpha\ dx$$

On intègre par parties

$$E(\ln(X))=\Big[\ln(x)\ x^{\alpha+1}\Big]_0^1-\int_0^1\frac{1}{x}\ x^{\alpha+1}\ dx=-\int_0^1 x^\alpha\ dx=\frac{-1}{\alpha+1}$$

D'après la loi forte des grands nombres, la moyenne empirique des $\ln(X_i)$ converge presque surement vers $\frac{-1}{\alpha+1}$, donc la suite des $\frac{-n}{\sum_{i=1}^n\ln(X_i)}$ converge presque surement vers $\alpha+1$. Et donc $\widehat{a_n}=\frac{-n}{\sum_{i=1}^n\ln(X_i)}-1$ est un estimateur fortement consistant de α.

6)

$$E((\ln(X))^2)=\int_0^1(\ln(x))^2\ (\alpha+1)\ x^\alpha\ dx=\Big[(\ln(x))^2\ x^{\alpha+1}\Big]_0^1-\int_0^1\frac{2\ln(x)}{x}\ x^{\alpha+1}\ dx=-2\int_0^1\ln(x)\ x^\alpha\ dx$$

donc $E((\ln(X))^2)=\frac{2}{(\alpha+1)^2}$ et $\mathrm{Var}(X)=\frac{1}{(\alpha+1)^2}$.
Le théorème central limite appliqué aux $Y_i=\ln(X_i)$ donne

$$P\left(\left|\sqrt{625}(\alpha+1)\left(\overline{Y_{625}}+\frac{1}{\alpha+1}\right)\right|\le2,05\right)\simeq P\left(|Z|\le2,05\right)\simeq96\%$$

qu'on peut réécrire sous la forme

$$P\left(25\left|(\alpha+1)\overline{Y_{625}}+1\right|\le2,05\right)\simeq96\%$$

$$P\left(\frac{-2,05}{25} \leq (\alpha+1)\overline{Y_{625}} + 1 \leq \frac{2,05}{25}\right) = P\left(\frac{-2,05/25-1}{\overline{Y_{625}}} \leq \alpha+1 \leq \frac{2,05/25-1}{\overline{Y_{625}}}\right) \simeq 96\%$$

On en conclut que

$$P(\alpha \in J) \simeq 96\% \quad \text{pour} \quad J = \left[-\frac{1}{\overline{Y_{625}}} \pm \frac{2,05}{25\ \overline{Y_{625}}} - 1\right]$$

En remarquant que $\frac{1}{\overline{Y_{625}}} = -\widehat{a_{625}} - 1$ l'intervalle de confiance à 96% obtenu peut aussi s'écrire

$$J = \left[\widehat{a_{625}} \pm \frac{2,05}{25}(\widehat{a_{625}} + 1)\right]$$

Université Lille1 Sciences et Technologies

U.F.R. de Mathématiques

MASS 3 S5 Estimation **Année 2014-2015 M.Fradon**

Examen, deuxième session, 2 mars 2015

Durée : 3 heures.

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

Ex 1. Meubles et poumons

Les ouvriers des usines de mobilier sont souvent exposés aux solvants et autres produits volatils, qui peuvent avoir un effet néfaste sur les poumons. L'un des indicateurs de la santé des poumons est le nombre maximal de litres d'air expiré par seconde (mesuré en demandant à la personne de souffler dans un tube relié à un debimètre).

1) On mesure ce nombre chez sept ouvriers. On note $X_1,\dots,X_7$ les résultats. Les X_i sont gaussiens. Leur espérance μ et leur écart-type σ sont des indicateurs de l'état de santé pulmonaire des ouvriers. Construire un intervalle de confiance pour l'estimation de μ, et un autre pour l'estimation de σ, au niveau de confiance 90% dans les deux cas.

2) Chez les sept ouvriers, le médecin du travail a observé les données suivantes :

$$3,94 \quad 1,47 \quad 2,06 \quad 2,36 \quad 3,74 \quad 3,43 \quad 3,79$$

Quelles sont les valeurs observées des intervalles de confiance ?

Ex 2. Dimensionnement d'un équipement téléphonique

Une société souhaite changer l'équipement téléphonique d'un site sur lequel travaillent 4900 employés. On sait (par les factures téléphoniques antérieures) qu'à chaque instant, la probabilité pour un employé d'être en train de téléphoner est de 7%. Le nouveau système téléphonique devra être dimensionné en conséquence. On note a le nombre total de lignes téléphoniques dont disposera le site. Dès que le nombre de personnes au téléphone dépassera a, il y aura saturation, c'est-à-dire que certaines communications ne pourront pas passer.

1) En utilisant le théorème central limite, déterminer quelle est la valeur minimale de a pour que la probabilité de saturation soit inférieure à 3%.

2) En réalité, dans les équipements téléphoniques, les lignes sont posées par groupes de trente. L'entreprise décide donc de s'équiper de 390 lignes en tout. Donner une valeur approximative de la probabilité de saturation. Calculer la précision de cette approximation. Peut-on réellement affirmer qu'avec 390 lignes, la probabilité de saturation ne dépasse pas 3% ?

Ex 3.

Le résultat X d'une expérience aléatoire a pour densité de probabilité la fonction f_α définie par $f_\alpha(x) = \dfrac{\alpha}{\sqrt{x}} e^{-2\alpha\sqrt{x}}$ si x est strictement positif et $f_\alpha(x) = 0$ sinon. Ici, α est un paramètre strictement positif fixé.

1) Quelle est la fonction de répartition F_α de la loi de probabilité de X ?

2) Prouver que X a une espérance et la calculer.

3) On veut simuler une variable aléatoire de même loi que X, donc de densité f_α. Décrire une méthode permettant de le faire.

4) Dans le cas où le paramètre α est inconnu, on cherche à l'estimer. On répète n fois l'expérience conduisant au tirage de X, de façon indépendante. On obtient ainsi les résultats $X_1, X_2, \ldots, X_n$. Déterminer l'estimateur $\hat{\alpha_n}$ du maximum de vraisemblance pour le paramètre α.

5) Déterminer la loi de la variable aléatoire $\sqrt{X}$.

6) $\hat{\alpha_n}$ est-il un estimateur fortement consistant de α ?

Ex 4. Loi de Poisson de grand paramètre

1) Les v.a. indépendantes X et Y suivent des lois de Poisson. La loi de X a pour paramètre α, celle de Y a le paramètre β. Prouver que $X + Y$ suit la loi de Poisson de paramètre $\alpha + \beta$.

2) Les v.a. X_1, X_2, X_3,... sont indépendantes, toutes de loi de Poisson de paramètre a. Déduire de la question précédente la loi suivie par $X_1 + \cdots + X_n$.

3) Calculer, pour n très grand, une valeur approximative de $P\left(\sum_{i=1}^{n} X_i \ \leq na\right)$.

4) Pour n assez grand, la loi de Poisson de paramètre na est proche d'une loi gaussienne. Indiquer laquelle.

5) Combien vaut $\lim_{n\to\infty} e^{-n} \sum_{k=0}^{n} \frac{n^k}{k!}$?

Université Lille1 Sciences et Technologies

U.F.R. de Mathématiques

MASS 3 S5 Estimation **Année 2014-2015 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 2014-2015 M.Fradon**

Devoir surveillé, 7 novembre 2014

Durée : 2 heures.

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

Ex 1. Dessiner le graphe de la fonction de répartition

$$\forall t \in \mathbb{R} \quad F(t) = \min(t^2, \frac{1}{4})\mathbf{1}_{0\leq t<1} + \mathbf{1}_{t\geq 1}$$

On veut réaliser des tirages de la loi correspondante à partir de tirages uniformes sur $]0;1[$. Que doit-on faire ?

Ex 2. Dératisation du campus

Un test biologique a une probabilité 3% de produire un "faux positif" : chez une souris qui n'a jamais eu de contact avec la maladie, le test a une probabilité $0,03$ d'être positif (c'est-à-dire d'affirmer que la souris est malade alors qu'elle ne l'est pas).

1) $n = 12$ souris sont élevées sur le campus par des biologistes. Ils veillent à ce que les cages, les instruments, la nourriture, etc, évitent aux souris tout contact avec la maladie. On fait subir le test à ces souris. En utilisant l'inégalité de Tchebychev, majorer la probabilité que 3 souris ou plus aient un test positif.

2) Sur les 12 souris élevées en cage, trois ont un test positif. Les biologistent réclament une campagne de dératisation du campus. Ils affirment que ces 25% de résultats positifs ne s'expliquent que par la mise en contact des souris de laboratoire avec la maladie. Ce qui prouve que des rongeurs sauvages malades ont pénétré dans le laboratoire. On leur répond que ces résultats positifs ont pu se produire par hasard, que l'événement constaté n'était pas si improbable...

Calculer combien de souris doivent élever les biologistes pour qu'il y ait moins d'une chance sur cent qu'un quart des souris (ou plus) aient un test positif, en l'absence de contamination.

Ex 3. Les lois de Pareto servent, entre autres, à modéliser la répartition des revenus. La loi de Pareto de paramètre $\alpha > 0$ est définie par sa densité sur $\mathbb{R}$:

$$f_\alpha \ : \ x \longmapsto \frac{\alpha}{x^{\alpha+1}}\mathbf{1}_{x>1}$$

On tire X_1, X_2, X_3,... indépendantes de loi de Pareto, avec un paramètre α fixé.

1) Pour quelle(s) valeurs de α la suite des moyennes empiriques $\overline{X}_n$ converge-t-elle presque surement, et quelle est alors sa limite ?

2) Pour quelle(s) valeurs de α le théorème central limite permet-il d'affirmer que la suite des moyennes empiriques $\overline{X_n}$ centrées normalisées converge en loi ?

3) Dans le cas où $\alpha = 4$, donner une valeur approximative de la probabilité que $\overline{X}_{400} > 1,4$

4) Pour quelle(s) valeurs de α le théorème de Berry-Esséen est-il utilisable ici ?

5) Dans le cas où $\alpha = 4$, calculer $\rho^3 = E\left(|X_1 - E(X_1)|^3\right)$. En déduire un encadrement de la valeur de $P(\overline{X}_{400} > 1,4)$.

Ex 4. Le graphique ci-dessous présente la répartition des tailles (en mm) des saumons pêchés dans un lac. Plus précisément, la courbe la plus à droite est la fonction de répartition empirique observée pour les tailles $X_1, \ldots, X_{168}$ des saumons mâles pêchés. La courbe à gauche est la fonction de répartition empirique observée pour les tailles $Y_1, \ldots, Y_{147}$ des saumons femelles.

1) Quel résultat mathématique assure que, pour un nombre de données suffisamment grand, la fonction de répartition empirique des X_i est proche de la fonction de répartition de la loi des X_i ?

2) On décide de modéliser la loi des X_i par la gausienne $\mathcal{N}(\mu_m, \sigma_m^2)$ et celle des Y_i par la gausienne $\mathcal{N}(\mu_f, \sigma_f^2)$. Au vu du graphique, semble-t-il plus raisonnable de choisir $\mu_m > \mu_f$ ou $\mu_m < \mu_f$? Justifier ce choix. De même, semble-t-il raisonnable de prendre $\sigma_m > \sigma_f$ ou $\sigma_m < \sigma_f$? Quel détail du graphique pousse à faire ce choix ?

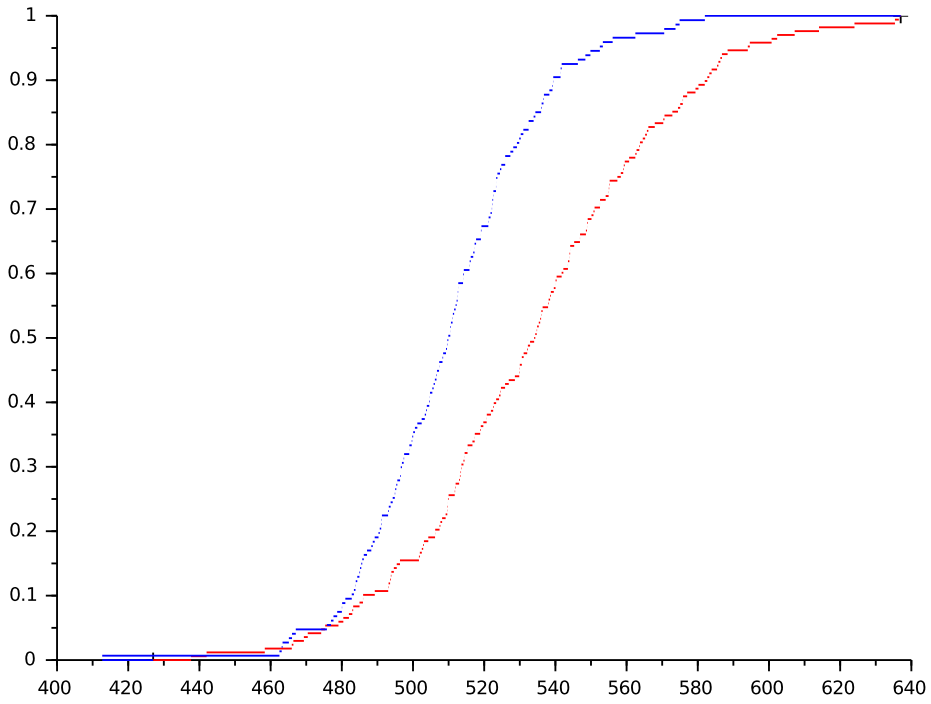

FIGURE 1 – saumons femelles (courbe à gauche) et mâles (courbe à droite)

Corrigé du devoir surveillé du 7 novembre 2014

Ex 1. La fonction de répartition de X est

$$\forall t \in \mathbb{R} \quad F(t) = t^2 \mathbf{1}_{0 \leq t < 1/2} + \frac{1}{4}\mathbf{1}_{1/2 \leq t < 1} + \mathbf{1}_{t \geq 1}$$

On a besoin de son pseudo-inverse.

$$\forall u \in]0; \frac{1}{4}] \quad F^{-1}(u) = \inf\{t \in \mathbb{R} \ ; \ F(t) \geq u\} = \inf\{t \geq 0 \ ; \ t^2 \geq u\} = \sqrt{u}$$

$$\forall u \in]\frac{1}{4}; 1[\quad F^{-1}(u) = \inf\{t \in \mathbb{R} \ ; \ F(t) \geq u\} = \inf\{t \geq 1 \ ; \ 1 \geq u\} = 1$$

On obtient la fonction quantile :

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

Si U est obtenu par tirage uniforme sur $]0; 1[$, alors $Y = \sqrt{U}\mathbf{1}_{0 < U \leq 1/4} + \mathbf{1}_{1/4 < U < 1}$ a même loi que X.

Ex 2. Dératisation du campus

1) Pour i de 1 à n, notons V_i l'indicatrice du fait que la souris numéro i a un test positif. Dans cette question $n = 12$. Pour des souris non contaminées, les V_i sont i.i.d. de loi $\mathcal{B}er(0,03)$. On doit calculer

$$P\left(\sum_{i=1}^{n} V_i \geq \frac{n}{4}\right) = P\left(\overline{V_n} \geq 0,25\right)$$

La proportion de souris contaminées ne peut pas être négative, donc l'événement $\left\{\overline{V_n} \geq 0,25\right\}$ est égal à

$$\left\{\overline{V_n} \leq -0,19 \text{ ou } \overline{V_n} \geq 0,25\right\} = \left\{\overline{V_n} - 0,03 \leq -0,22 \text{ ou } \overline{V_n} - 0,03 \geq 0,22\right\} = \left\{\left|\overline{V_n} - 0,03\right| \geq 0,22\right\}$$

L'inégalité de Tchebychev donne

$$P\left(\sum_{i=1}^{n} V_i \geq \frac{n}{4}\right) = P\left(\left|\overline{V_n} - 0,03\right| \geq 0,22\right) \leq \frac{0,03(1-0,03)}{n(0,22)^2} \simeq 0,0501$$

La probabilité que trois souris ou plus aient un test positif est inférieure à $5,01\%$.

2) On veut trouver un nombre de souris n pour lequel $P\left(\sum_{i=1}^n V_i \geq \frac{n}{4}\right) \leq 1\%$. Le calcul a été fait à la question précédente pour un entier n quelconque. Il suffit donc ici de trouver n tel que le majorant $\frac{0{,}03(1-0{,}03)}{n(0{,}22)^2}$ soit inférieur à $0{,}01$.

$$\frac{0,03(1-0,03)}{n(0,22)^2} \leq 0,01 \quad \Longleftrightarrow \quad n \geq \frac{0,03(1-0,03)}{0,01(0,22)^2} = \frac{300(100-3)}{22^2} \simeq 60,124$$

Les biologistes doivent donc élever au moins 61 souris avec du matériel non contaminé. Si, comme lors du premier élevage, un quart d'entre elles sont testées positives à la maladie, il pourront raisonnablement réclamer une campagne de dératisation du campus [1].

Ex 3. 1) Les X_i sont i.i.d. de loi Pareto(α) avec $\alpha > 0$ fixé. La loi forte des grands nombres affirme que $\overline{X}_n$ converge presque surement si et seulement si cette loi a une espérance. Il faut donc vérifier l'intégrabilité des X_i.

$$E(|X_1|) = \int_{\mathbb{R}} |x| f_\alpha(x)\, d\lambda(x) = \int_{\mathbb{R}} |x| \frac{\alpha}{x^{\alpha+1}} \mathbf{1}_{x>1}\, d\lambda(x) = \alpha \int_{[1;+\infty[} \frac{1}{x^\alpha}\, d\lambda(x)$$

Cette intégrale est convergente si $\alpha > 1$, elle vaut $+\infty$ si $0 < \alpha \leq 1$. Donc $\overline{X}_n$ converge presque surement si et seulement si $\alpha > 1$. Dans ce cas sa limite vaut

$$E(X_1) = \int_{\mathbb{R}} x f_\alpha(x)\, d\lambda(x) = \alpha \int_{[1;+\infty[} \frac{1}{x^\alpha}\, d\lambda(x) = \alpha \int_1^{+\infty} x^{-\alpha}\, dx = \alpha \left[\frac{x^{-\alpha+1}}{-\alpha+1}\right]_1^{+\infty} = \frac{\alpha}{\alpha-1}$$

2) Le théorème central limite donne la convergence en loi de la suite des moyennes empiriques centrées normalisées dès lors que la variance est strictement positive. Il faut trouver pour quels α la loi Pareto(α) a un moment d'ordre deux, et voir si la variance est non-nulle.

$$E((X_1)^2) = \int_{\mathbb{R}} x^2 \frac{\alpha}{x^{\alpha+1}} \mathbf{1}_{x>1}\, d\lambda(x) = \alpha \int_{[1;+\infty[} \frac{1}{x^{\alpha-1}}\, d\lambda(x) < +\infty \quad \Longleftrightarrow \quad \alpha > 2$$

Et si $\alpha > 2$:

$$E((X_1)^2) = \alpha \int_1^{+\infty} x^{-\alpha+1}\, dx = \alpha \left[\frac{x^{-\alpha+2}}{-\alpha+2}\right]_1^{+\infty} = \frac{\alpha}{\alpha-2}$$

$$\mathrm{Var}(X_1) = \frac{\alpha}{\alpha-2} - \frac{\alpha^2}{(\alpha-1)^2} = \frac{\alpha}{(\alpha-1)^2(\alpha-2)} > 0$$

La condition pour pouvoir appliquer le théorème central limite à la loi Pareto(α) est donc $\alpha > 2$.

3) Dans le cas où $\alpha = 4$, les X_i sont i.i.d. d'espérance $\frac{4}{3}$ et de variance $\frac{2}{9}$. Le théorème central limite assure que $\sqrt{n}\frac{\overline{X_n}-\frac{4}{3}}{\sqrt{\frac{2}{9}}} = \sqrt{n}\frac{3\overline{X_n}-4}{\sqrt{2}}$ converge en loi vers $Z \sim \mathcal{N}(0;1)$. Pour $n = 400$

$$P\left(\overline{X_{400}} > 1,4\right) = P\left(\sqrt{400}\ \frac{3\overline{X_{400}}-4}{\sqrt{2}} > \sqrt{400}\ \frac{3\times 1,4-4}{\sqrt{2}}\right) \simeq P\left(Z > 20\ \frac{0,2}{\sqrt{2}}\right)$$

1. Pour ceux qui souhaitent connaitre la fin de l'histoire : les souris de laboratoire étaient bel et bien contaminées par des rongeurs sauvages. Les biologistes dont les recherches étaient perturbées ont obtenu la dératisation du campus.

On utilise la table gaussienne :

$$P\left(\overline{X_{400}} > 1,4\right) \simeq P\left(Z > 2,83\right) \simeq 1 - 0,9977 = 0,23\%$$

4) Pour utiliser le théorème de Berry-Esséen, on a besoin d'un moment d'ordre 3.

$$E(|X_1|^3) = \int_{\mathbb{R}} |x|^3 \frac{\alpha}{x^{\alpha+1}} \mathbf{1}_{x>1} \, d\lambda(x) = \alpha \int_{[1;+\infty[} \frac{1}{x^{\alpha-2}} \, d\lambda(x) < +\infty \quad \Longleftrightarrow \quad \alpha > 3$$

Remarquons que $\rho^3 = E(|X_1 - E(X_1)|^3)$ est strictement positif dès lors que $\alpha > 3$, car la variance est stritement positive :

$$0 < P(X_1 \neq E(X_1)) = \lim_{k\to+\infty} P(|X_1 - E(X_1)| > 1/k) \leq \lim_{k\to+\infty} k^3 \; E(|X_1 - E(X_1)|^3)$$

5) Dans le cas où $\alpha = 4$, calculons le moment centré d'ordre 3.

$$\rho^3 = E(|X_1 - E(X_1)|^3) = \int_{\mathbb{R}} |x - \frac{4}{3}|^3 \frac{4}{x^5} \mathbf{1}_{x>1} \, d\lambda(x) = 4\int_1^{4/3} (\frac{4}{3} - x)^3 \, \frac{1}{x^5} \, dx + 4\int_{4/3}^{+\infty} (x - \frac{4}{3})^3 \, \frac{1}{x^5} \, dx$$

Pour simplifier, on note $m = \frac{4}{3}$ l'espérance de la Pareto. Le premier terme vaut

$$\begin{aligned} 4\int_1^m \left(m^3 - 3m^2x + 3mx^2 - x^3\right) \frac{1}{x^5} \, dx &= 4\left[-m^3 \, \frac{x^{-4}}{4} + 3m^2 \, \frac{x^{-3}}{3} - 3m\frac{x^{-2}}{2} + x^{-1}\right]_1^m \\ &= \left[-\frac{m^3}{x^4} + \frac{4m^2}{x^3} - \frac{6m}{x^2} + \frac{4}{x}\right]_1^m = \frac{1}{m} - \left(-m^3 + 4m^2 - 6m + 4\right) \end{aligned}$$

et le deuxième est égal à

$$4\int_m^{+\infty} \left(-m^3 + 3m^2x - 3mx^2 + x^3\right) \, \frac{1}{x^5} \, dx = -\left[-\frac{m^3}{x^4} + \frac{4m^2}{x^3} - \frac{6m}{x^2} + \frac{4}{x}\right]_m^{+\infty} = \frac{1}{m}$$

donc

$$\rho^3 = \frac{2}{m} - \left(-m^3 + 4m^2 - 6m + 4\right) = \frac{3}{2} + \frac{64}{27} - \frac{64}{9} + 8 - 4 = \frac{11}{2} - \frac{128}{27} = \frac{41}{54} \simeq 0,7593$$

Le théorème de Berry-Esséen donne

$$\left|P\left(\overline{X_{400}} > 1,4\right) - P\left(Z > 2,83\right)\right| \leq \frac{\frac{41}{54}}{2\sqrt{400}(\frac{2}{9})^{3/2}} = \frac{41\sqrt{2}}{320} \simeq 0,1812$$

Finalement, $P\left(\overline{X_{400}} > 1,4\right)$ est égale à $0,0023$ à $0,1812$ près. Autrement dit $P\left(\overline{X_{400}} > 1,4\right)$ est entre 0% et $18,35\%$. On est ici dans l'un des cas où le théorème central limite donne un résultat imprécis. Le théorème de Berry-Esséen nous alerte sur cette imprécision.

Ex 4. 1) Le théorème de Glivenko-Cantelli donne la convergence uniforme presque sûre de la fonction de répartition empirique $t \longmapsto \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{X_i \leq t}$ vers la la fonction de répartition F_X de la loi des X_i. De même pour les Y_i. Les données dont on dispose ici (168 pour les saumons mâles et 147 pour les femelles) sont assez nombreuses pour voir se dessiner approximativement F_X et F_Y sur le graphique.

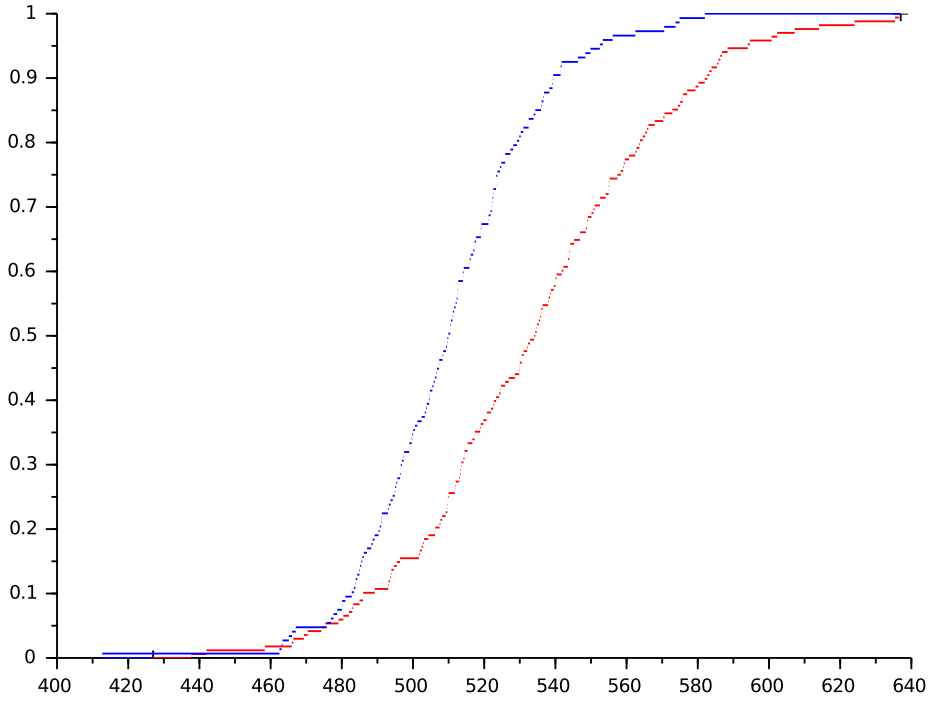

FIGURE 2 – saumons femelles (courbe à gauche) et mâles (courbe à droite)

2) On décide d'utiliser la modélisation $X_i \sim \mathcal{N}(\mu_m, \sigma_m^2)$ et $Y_i \sim \mathcal{N}(\mu_f, \sigma_f^2)$. Avec cette modélisation $F_X(\mu_m) = 0,5$ et $F_Y(\mu_f) = 0,5$. Au vu des versions approchées de F_X et F_Y dessinées sur le graphique, on est amené à penser que $F_X^{-1}(0,5) > F_Y^{-1}(0,5)$ (la courbe correspondant aux mâles est assez nettement à droite de celle pour les femelles). Il est donc raisonnable de modéliser avec $\mu_m > \mu_f$.

La pente de la fonction de répartition d'une gaussienne augmente quand son écart-type diminue. Autrement dit, $\sigma_m < \sigma_f$ si F_X "monte plus vite" que F_Y et $\sigma_m > \sigma_f$ sinon. Sur le graphique, la courbe de gauche (approximation de F_Y) présente une pente plus raide que celle de droite (approximation de F_X). Le plus raisonnable est donc de choisir $\sigma_m > \sigma_f$.