**Université Pierre et Marie Curie** **Année 2011–2012**
Licence - Mathématiques - LM346 Processus et simulations

**Corrigé de la série d'exercices N° 2**

**Exercice 1. Quantiles**
Puisque $-X$ a la même loi que $X$, on a, pour tout réel $a$ :

$$P(X < -a) = P(-X < -a) = P(X > a)$$

Calculons alors :

$$\begin{aligned} P(|X| \leq q_x) &= P(-q_x \leq X \leq q_x) \\ &= 1 - P(X \in ]-\infty, -q_x[ \cup ]q_x, +\infty[) \\ &= 1 - (P(X < -q_x) + P(X > q_x)) \text{ car ces deux événements sont disjoints} \\ &= 1 - 2P(X > q_x) = 1 - 2x \text{ d'après ce qui précède.} \end{aligned}$$

Notons qu'on n'a finalement pas utilisé l'hypothèse de continuité de la fonction de répartion de $X$ à part pour justifier l'existence de $q_x$ pour tout $x$.

**Exercice 2. Intervalle de confiance pour la méthode de Monte-Carlo**
Commençons par remarquer que l'exercice précédent s'applique, en particulier, à la loi normale. Par conséquent, $\int_{-q_{1/200}}^{q_{1/200}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx P(|N| \leq q_{1/200}) = 1 - \frac{2}{200} = 99\%$.
On note $I(f) = \int_{[0,1]^d} f(x)dx = E(f(X_1))$ et $\sigma = (Var(f(X_1)))^{1/2}$. Appliquons le théorème 3.2.2 à la suite $(X_i)_{i\geq 1}$ :

$$\left| P\left( |\frac{1}{n} S_n - I(f)| < \frac{\sigma q_{1/200}}{\sqrt{n}} \right) - 99\% \right| \leq 6\big(\frac{||f||_\infty}{\sigma}\big)^3 \frac{1}{\sqrt{n}}$$

Si $n$ est tel que $6\big(\frac{||f||_\infty}{\sigma}\big)^3 \frac{1}{\sqrt{n}} \leq 1\%$, on a donc

$$P\left( |\frac{1}{n} S_n - I(f)| < \frac{\sigma q_{1/200}}{\sqrt{n}} \right) \geq 99\% - 1\% = 98\%$$

Si en outre $\frac{\sigma q_{1/200}}{\sqrt{n}} \leq \frac{1}{1000}$, alors

$$P\left( I(f) \in [\frac{S_n}{n} - 1‰, \frac{S_n}{n} + 1‰] \right) \geq P\left( |\frac{1}{n} S_n - I(f)| < \frac{\sigma q_{1/200}}{\sqrt{n}} \right) \geq 98\%$$

Il suffit donc de prendre $n_0 \geq \sup\Big(\sigma^2 q_{1/200}^2 10^6,\, 6(||f||_\infty/\sigma)^6 10^4\Big)$.

**Exercice 3. Intervalle de confiance non asymptotique pour le paramètre d'une loi de Bernoulli**
On note $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.
1. D'après l'inégalité de Tchebychev,

$$\begin{aligned} P(|\frac{1}{n}\sum_{i=1}^n X_i - \theta| \geq \varepsilon) &\leq \frac{1}{\varepsilon^2} Var\left( \frac{1}{n}\sum_{i=1}^n (X_i - \theta) \right) \\ &\leq \frac{1}{n^2\varepsilon^2} \sum_{i=1}^n Var(X_i - \theta) \text{ par indépendance des } X_i \\ &\leq \frac{\theta(1-\theta)}{n\varepsilon^2} \text{ car } Var(X_i - \theta) = Var(X_1) = \theta(1-\theta) \end{aligned}$$

On vérifie que, pour tout $p \in \mathbb{R}$, $p(1-p) \leq \frac{1}{4}$. En effet, la fonction $x \mapsto x(1-x)$ est un polynôme du second degré, qui admet un extremum au point où sa dérivée $x \mapsto 1-2x$ s'annule, donc en $1/2$. Cet extremum est un maximum puisque le terme de degré deux a un coefficient négatif. Par conséquent, pour tout réel $\theta$, $\theta(1-\theta) \leq 1/2(1-1/2) = 1/4$. On obtient le résultat en combinant cette majoration avec l'inégalité précédente.

2. Soit $\alpha \in ]0,1[$. Posons $\varepsilon = \frac{1}{2\sqrt{n\alpha}}$. D'après la première question,

$$P(|\theta - \overline{X}_n| \geq \varepsilon) \leq \alpha \text{ i.e. } P(\overline{X}_n - \varepsilon \leq \theta \leq \overline{X}_n + \varepsilon) \leq \alpha,$$

ce qui revient à dire que l'intervalle $[\overline{X}_n - \varepsilon, \overline{X}_n + \varepsilon]$ définit un intervalle de confiance de fiabilité au moins $100(1-\alpha)\%$ pour l'estimation de $\theta$.
3. Posons $X_i = 0$ si la $i$-ème personne répond non, et $X_i = 1$ si elle répond oui.
On fait l'hypothèse que les $X_i$ sont i.i.d.; elles suivent alors une loi de Bernoulli de paramètre commun $\theta$.

**Exercice 4. : Intervalles de confiance non asymptotiques pour les paramètres de lois gaussiennes.**
1. La variable $n\bar{Y}_n$ est la somme de $n$ gaussiennes indépendantes de loi $\mathcal{N}(\mu, \sigma^2)$, donc elle suit une loi gaussienne $\mathcal{N}(n\mu, n\sigma^2)$. Donc $\bar{Y}_n$ est une gaussienne de moyenne $\mu$ et de variance $\frac{\sigma^2}{n}$, et par suite,

$$\frac{\sqrt{n}(\bar{Y}_n - \mu}{\sigma} \sim \mathcal{N}(0,1).$$

2. D'après la définition des quantiles de la loi $\mathcal{N}(0,1)$, on a pour $x < \frac{1}{2}$ :

$$P\left(\left|\frac{\sqrt{n}(\bar{Y}_n - \mu)}{\sigma}\right| \leq q_x\right) = 1 - 2x,$$

car la loi $\mathcal{N}(0,1)$ est symétrique. Cela peut se réécrire de la manière suivante :

$$P\left(\bar{Y}_n - \frac{q_x\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y}_n + \frac{q_x\sigma}{\sqrt{n}}\right) = 1 - 2x.$$

En prenant $x = \frac{\alpha}{2}$, on obtient un intervalle de confiance de fiabilité $100(1-\alpha)\%$ pour $\mu$ si l'on connaît $\sigma$.

3. Si l'on essaie de faire exactement la même chose pour $\sigma$, on ne trouvera qu'une borne inférieure pour $\sigma$. En revanche, en prenant $x < y < \frac{1}{2}$, on peut écrire :

$$P\left(q_y < \frac{\sqrt{n}|\bar{Y}_n - \mu|}{\sigma} \leq q_x\right) = P\left(\left|\frac{\sqrt{n}(\bar{Y}_n - \mu}{\sigma}\right| \leq q_x\right) - P\left(\left|\frac{\sqrt{n}(\bar{Y}_n - \mu}{\sigma}\right| \leq q_y\right) = 2(y-x).$$

Cela peut être réécrit de la manière suivante :

$$P\left(\frac{\sqrt{n}}{q_x}|\bar{Y}_n - \mu| \leq \sigma < \frac{\sqrt{n}}{q_y}|\bar{Y}_n - \mu|\right) = 2(y-x).$$

En choisissant $y$ et $x$ de sorte que $y - x = \frac{1-\alpha}{2}$, on obtient un intervalle de confiance de fiabilité $100(1-\alpha)\%$ pour $\sigma$ si on connaît $\mu$. On peut souhaiter prendre $x$ et $y$ qui minimisent la taille de l'intervalle dans lequel $\sigma$ va se trouver avec une grande probabilité.

La différence de forme de l'intervalle de confiance pour $\mu$ et $\sigma$ peut s'expliquer de la manière suivante : la moyenne est le "milieu" de la densité pour une gaussienne, il faut donc regarder la

partie centrale de la distribution (entre $-q_x$ et $q_x$). L'écart-type représente quant à lui la "largeur à mi-hauteur" de la gaussienne. Pour l'estimer, ce qui va compter ce sont les parties de la distribution ni trop près, ni trop loin du centre, c'est à dire entre $-q_y$ et $-q_x$ d'une part, et entre $q_x$ et $q_y$ d'autre part.

Comme la gaussienne est une distribution à densité, on peut remplacer les inégalités strictes par des inégalités larges dans les probabilités sans rien changer au résultat.

4. D'après le cours (théorème 4.2.1), la variable $\frac{\bar{Y}_n-\mu}{\sqrt{V_n/(n-1)}}$ suit une loi de Student à $n-1$ degrés de liberté.

5. Si les $q_x$ sont maintenant les quantiles de la loi de Student à $n-1$ degrés de liberté (qui dépendent de $n$ donc !!), on peut écrire, comme les lois de Student sont symétriques, que

$$P\left(-q_x \leq \frac{\bar{Y}_n - \mu}{\sqrt{V_n/(n-1)}} \leq q_x\right) = 1 - 2x.$$

ce qui se réécrit sous la forme suivante :

$$P\left(\bar{Y}_n - \sqrt{\frac{V_n}{n-1}} \leq \mu \leq \bar{Y}_n + \sqrt{\frac{V_n}{n-1}}\right) = 1 - 2x.$$

Cette dernière expression donne donc un intervalle de confiance de fiabilité $100(1-\alpha)\%$ pour $\mu$ lorsque $\sigma$ est inconnu, en choisissant $x = \frac{\alpha}{2}$.

6. D'après le cours, la variable $\frac{nV_n}{\sigma^2}$ suit une loi du $\chi^2$ à $n-1$ degrés de liberté.

7. On est dans une situation similaire à la question 3, òu l'on cherche un intervalle de confiance pour une quantité positive à partir d'une distribution non symétrique. Soit $0 < x < y < 1$. Si $q_x$ et $q_y$ dénotent les quantiles de niveau $x$ et $y$ respectivement de la loi du $\chi^2$ à $n-1$ degrés de liberté (qui dépendent donc de $n$), on peut écrire :

$$P\left(q_y < \frac{nV_n}{\sigma} \leq q_x\right) = P\left(\frac{nV_n}{\sigma} \leq q_x) - P(\frac{nV_n}{\sigma} \leq q_y\right) = y - x,$$

c'est-à-dire :

$$P\left(\frac{nV_n}{q_x} \leq \sigma < \frac{nV_n}{q_y}\right) = y - x.$$

En choisissant $y$ et $x$ de sorte que $y - x = 1 - \alpha$, on obtient un intervalle de confiance pour $\sigma$ de fiabilité $100(1-\alpha)\%$ lorsque $\mu$ est inconnu. Parmi tous les choix possibles, on peut vouloir prendre celui qui minimise la taille de l'intervalle, c'est à dire qui minimise $\frac{1}{q_x} - \frac{1}{q_y}$, à $y - x$ fixé.

**Exercice 5. Intervalles de confiance asymptotiques pour une suite de variables aléatoires i.i.d.**

1. Comme les $Y_i$ sont dans $L^2$, donc dans $L^1$, d'après la loi (forte) des grands nombres, $\bar{Y}_n$ converge presque-sûrement vers $E(Y_1) = \mu$. D'autres part, d'après le théorème central limite, comme les variables sont de carré intégrable, $\frac{\sqrt{n}(\bar{Y}_n-\mu)}{\sigma}$ converge en loi vers une loi normale $\mathcal{N}(0,1)$.

On a donc que

$$\lim_{n\to+\infty} P(\bar{Y}_n - q_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y}_n + q_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

ce qui est un intervalle de confiance asymptotique pour $\mu$ de fiabilité $100(1-\alpha)\%$.

2. Il s'agit d'un calcul analogue à celui pour montrer que les deux formules $E((X-E(X))^2)$ et $E(X^2)-E(X)^2$ pour la variance d'une variable aléatoire $X$ de carré intégrables sont égales.

$$\begin{aligned}
V_n &= \frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i^2-2\bar{Y}_nY_i+\bar{Y}_n^2)\\
&= \frac{1}{n}\sum_{i=1}^{n}Y_i^2 - 2\frac{\bar{Y}_n}{n}\sum_{i=1}^{n}Y_i + \frac{1}{n}\sum_{i=1}^{n}\bar{Y}_n^2\\
&= \left(\frac{1}{n}\sum_{i=1}^{n}Y_i^2\right) - 2\bar{Y}_n^2 + \bar{Y}_n^2\\
&= \frac{1}{n}\sum_{i=1}^{n}Y_i^2 - \bar{Y}_n^2 = \frac{1}{n}\sum_{i=1}^{n}Y_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}Y_i\right)^2
\end{aligned}$$

Les $Y_i^2$ sont des variables intégrables i.i.d. D'après la loi forte des grands nombres, la première somme $\frac{1}{n}\sum_{i=1}^{n}Y_i^2$ converge vers $E(Y_1^2)$ presque-sûrement. Comme $\bar{Y}_n$ converge presque-sûrement vers $E(Y_1)$, on en déduit que $V_n$ converge presque-sûrement vers $E(Y_1^2)-E(Y_1)^2 = Var(Y_1) = \sigma^2$.

3. Si l'on admet que $\frac{\sqrt{n}(\bar{Y}_n-\mu)}{\sqrt{V_n}}$ converge en distribution vers une loi $\mathcal{N}(0,1)$, on a par le même argument qu'à la première question :

$$\lim_{n\to+\infty} P(\bar{Y}_n - q_{\alpha/2}\frac{\sqrt{V_n}}{\sqrt{n}} \leq \mu \leq \bar{Y}_n + q_{\alpha/2}\frac{\sqrt{V_n}}{\sqrt{n}}) = 1-\alpha$$

ce qui est un intervalle de confiance asymptotique pour $\mu$ de fiabilité $100(1-\alpha)\%$, lorsque $\sigma$ est inconnu.

4. Pour cette appliquation numérique, on a $n=100$, $\bar{Y}_n = 175$, $V_n = 49$, $\alpha = 0,05$ et $q_{\alpha/2} = 1,96$. On trouve ainsi qu'avec une fiabilité de 95%, la taille moyenne de la population est comprise entre $173,628$ cm et $176,372$cm .

**Exercice 6. Intervalles de confiance asymptotiques pour une suite de vaiid de Bernouilli**
1. Les loi de Bernouilli sont de carré intégrable (car bornées). On est donc dans le cadre du théorème central limite. Comme $E(Y_1) = p$ et $\mathrm{Var}(Y_1) = p(1-p)$, on a que

$$\frac{\sqrt{n}(\bar{Y}_n - p)}{p(1-p)}$$

converge en loi vers une loi normale centrée réduite $\mathcal{N}(0,1)$. Lorsque $q$ est égal à $q_{\alpha/2}$ le quantile de niveau $\alpha/2$ de la loi $\mathcal{N}(0,1)$, on a le résultat de convergence suivant :

$$\lim_{n\to+\infty} P\left(p \in \left[\bar{Y}_n - \frac{q\sqrt{\bar{Y}_n(1-\bar{Y}_n)}}{\sqrt{n}}, \bar{Y}_n + \frac{q\sqrt{\bar{Y}_n(1-\bar{Y}_n)}}{\sqrt{n}}\right]\right) = 1-\alpha$$

(en admettant la même chose que dans la question 3 de l'exercice précédent).

2. En remplaçant $\bar{Y}_n(1-\bar{Y}_n)$ par $\frac{1}{4}$, on agrandit l'intervalle dans lequel on autorise $p$ donc on augmente la probabilité :

$$\lim_{n\to+\infty} P\left(p \in \left[\bar{Y}_n - \frac{q}{2\sqrt{n}}, \bar{Y}_n + \frac{q}{2\sqrt{n}}\right]\right) \geq 1-\alpha.$$

Il s'agit donc d'un intervalle de confiance pour $p$ de fiabilité au moins $100(1-\alpha)\%$.

4. On a $n = 1000$ et $\bar{Y}_n = \frac{250}{1000} = \frac{1}{4}$ (proportion des gens sondés à répondre "oui"). Pour avoir une fiabilité supérieure à 95%, on prend $q = 1,96$, et on obtient qu'avec une probabilité supérieure à 95%, la proportion de Français qui auraient répondu "oui" était comprise entre $21,9\%$ et $28,1\%$.

5. On a $n = 900$ et $\bar{Y}_n = \frac{639}{900} = 0,71$. Lorsque $q = 1,96$, on obtient qu'avec une probabilité supérieure à 95%, la proportion de gens aux États-Unis croyant au diable est comprise entre $67,73\%$ et $74,27$. L'intervalle $[67\%, 75\%]$ correspond à une valeur de $q$ égale à $2 \times 0,04\sqrt{n} = 2,4$, ce qui correspond à une fiabilité supérieure à 99%.

**Exercice 7. Test de $\chi$-deux**

a) C'est la loi de $\chi$-deux à $r-1$ degrés de liberté, cad la loi de $X_1^2 + \cdots X_{r-1}^2$, $X_1, \ldots, X_r$ étant des variables aléatoires indépendantes d'espéranec 0 et de variance 1.

b) Soient $X_1, \ldots, X_{1000}$ des variables aléatoires indépendantes, de même loi, $X_k = i$ si le $k$ème sondé préfère le $i$ème forfait, $i = 1, 2, 3, 4, 5$. On veut tester l'hypothèse que $P(X_k = i) = p_i$ avec $p_1 = 0.1$, $p_2 = 0.2$, $p_3 = 0.3$, $p_4 = 0.3$, $p_5 = 0.1$. Alors, si l'hypothèse est correcte, $T_{1000}$ suit (approximativement !) la loi de $\chi$-deux à $4 = 5 - 1$ degrés de liberté.

Notons $\chi^2(4)$ une v.a. de cette loi. On trouve le nombre $\chi_{0.05,4}$ tel que $P(\chi^2(4) > \chi_{0.05,4}) = 0.05$. On a $P(\chi^2(4) < \chi_{0.05,4}) = 0.95$. Alors $\chi_{0.05,4} = x_{0.95,4}$ dans la table de la page web indiquée et $x_{0.95,4} = 9.49$. Si l'hypohèse est correcte, on a $P(T_{1000} > \chi_{0.05,4}) = 0.05$, cad l'événement $T_{1000} > \chi_{0.05,4}$ est trop peu probable, presque impossible.

La réalisation de $T_{1000}$ dans le sondage nous donne le nombre $T_{1000} = \frac{(80-100)^2}{100} + \frac{(220-200)^2}{200} + \frac{(330-300)^2}{300} + \frac{(280-300)^2}{300} + \frac{(90-100)^2}{100}$ (on laisse le lecteur le calculer). On compare ce nombre à celui $\chi_{0.05,4} = 9.49$.

Si on observe en réalité $T_{1000} > \chi_{0.05,4}$, comme la réalisation cet événement est trop peu probable sous notre hypothèse, on rejete l'hypothèse. Si on observe $T_{1000} \leq \chi_{0.05,4}$, on accepte l'hypothèse.

c) On calcule le nombre $T_{1000} = \frac{(149-150)^2}{150} + \frac{(202-200)^2}{200} + \frac{(53-50)^2}{50} + \frac{(96-100)^2}{100} + \frac{(301-300)^2}{300} + \frac{(199-200)^2}{200}$. Notons $\chi^2(5)$ une v.a. de loi $\chi$-2 à $6 - 1 = 5$ degrés de liberté. On trouve le nombre $\chi_{0.01,5}$ tel que $P(\chi^2(5) > \chi_{0.01,5}) = 0.01$. On a $P(\chi^2(5) < \chi_{0.01,5}) = 0.99$ Ce nombre $\chi_{0.01,5} = x_{0.99,5} = 15.09$ sur la page web indiquée.

Si on observe en réalité $T_{1000} > \chi_{0.05,5} = 15.09$, comme la réalisation cet événement est trop peu probable sous notre hypothèse, on rejete l'hypothèse. Si on observe $T_{1000} \leq \chi_{0.05,5}$, on accepte l'hypothèse.