Statistiques appliquées L3
Année universitaire 2006/2007

Corrigé de l'interrogation n°1

Exercice 1 (4 **points)**
1- $f_{X,Y}$ est une fonction de densité sur R^2 donc la valeur de k doit être telle que

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)dxdy = 1$$

Or $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)dxdy = \int_0^2\int_0^1 kx^3ydxdy = k\int_0^2 y\left(\int_0^1 x^3dx\right)dy = k\int_0^2 y\left[\frac{x^4}{4}\right]_0^1 dy$
$= \frac{k}{4}\int_0^2 ydy = \frac{k}{4}\left[\frac{y^2}{2}\right]_0^2 = \frac{k}{2}$ donc $k = 2$.
2- Pour tout $x \in R$, on a : $f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dy$. Deux cas doivent être distingués :
- Si $x \notin [0,1]$ alors $f_{X,Y}(x,y) = 0$ pour tout $y \in R$ d'où $f_X(x) = 0$.
- Si $x \in [0,1]$ alors $f_{X,Y}(x,y) = 2x^3y$ si $y \in [0,2]$ et $f_{X,Y}(x,y) = 0$ si $y \notin [0,2]$ d'où
$f_X(x) = \int_0^2 2x^3ydy = 2x^3\left[\frac{y^2}{2}\right]_0^2 = 4x^3$
Conclusion :

$$f_X(x) = \begin{cases} 4x^3 & \text{si} \quad 0 \le x \le 1 \\ 0 & \text{sinon} \end{cases}$$

Par un raisonnement similaire, on démontre que :

$$f_Y(y) = \begin{cases} \frac{1}{2}y & \text{si} \quad 0 \le y \le 2 \\ 0 & \text{sinon} \end{cases}$$

Il s'avère que pour tout $(x,y) \in [0,1]\times[0,2]$: $f_X(x)f_Y(y) = 2x^3y = f_{X,Y}(x,y)$
et pour tout $(x,y) \notin [0,1]\times[0,2]$: $f_X(x)f_Y(y) = 0 = f_{X,Y}(x,y)$, donc

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \quad \text{pour tout } (x,y) \in R^2$$

ce qui veut dire que les variables X et Y sont indépendantes.

Exercice 2 (4 **points) :**
Notons D l'évènement : "le bien est défectueux"
U_i l'évènement : "le bien provient de l'usine U_i" pour tout $i \in \{1,2,3\}$.
La probabilité qu'un bien choisi au hasard, et dont on constate qu'il est défectueux, provienne de l'usine U_i où $i \in \{1,2,3\}$ est donnée par

$$P(U_i \mid D) = \frac{P(U_i \cap D)}{P(D)} = \frac{P(U_i)P(D \mid U_i)}{P(U_1)P(D \mid U_1) + P(U_2)P(D \mid U_2) + P(U_3)P(D \mid U_3)}$$

La première égalité s'obtient par l'application de la définition de la probabilité conditionnelle $P(U_i \mid D)$. La seconde inégalité s'obtient par l'application de la définition de la probabilité conditionnelle $P(D \mid U_i)$ au numérateur et de la formule de la probabilité totale au dénominateur. Le tout forme la formule de Bayes.
Avec $P(U_1) = 0,3$, $P(U_2) = 0,3$, $P(U_3) = 0,4$ et $P(D \mid U_1) = 0,04$, $P(D \mid U_2) = 0,03$, $P(D \mid U_3) = 0,02$, on trouve :

$$P(U_1 \mid D) = \frac{12}{29};\ P(U_2 \mid D) = \frac{9}{29};\ P(U_3 \mid D) = \frac{8}{29}$$

Exercice 3 (12 points) :

1- Par linéarité de l'espérance :

$$E\left(\bar{X}_n\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{nm}{n} = m$$

$V(\bar{X}_n) = \frac{1}{n^2} V\left(\sum_{i=1}^{n} X_i\right)$ or les X_i sont indépendantes donc $V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i)$ d'où :

$$V(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Puisque $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ est une combinaison linéaire de variables indépendantes suivant une loi normale alors $\bar{X}_n$ suit également une loi normale. Les paramètres de cette loi sont $E\left(\bar{X}_n\right) = m$ et $V(\bar{X}_n) = \frac{\sigma^2}{n}$:

$$\bar{X}_n \rightsquigarrow N\left(m, \frac{\sigma^2}{n}\right)$$

2- Les X_i étant des variables indépendantes admettant la même espérance et la même variance, la loi faible des grands nombres permet d'affirmer directement que $\left(\bar{X}_n\right)$ converge en probabilité vers l'espérance commune des X_i c'est-à-dire m. De plus, les X_i suivent la même loi donc la loi forte des grands nombres est également applicable: $\left(\bar{X}_n\right)$ converge presque sûrement vers m.

3-a- L'application de l'inégalité de Tchebychev à $\bar{X}_n$ donne, pour tout $\varepsilon > 0$ fixé,

$$P(\left|\bar{X}_n - E(\bar{X}_n)\right| \geq \varepsilon) \leq \frac{V(\bar{X}_n)}{\varepsilon^2}$$

donc sachant que $E\left(\bar{X}_n\right) = m$ et $V(\bar{X}_n) = \frac{\sigma^2}{n}$, on obtient:

$$P(\left|\bar{X}_n - m\right| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \underset{n \longrightarrow +\infty}{\longrightarrow} 0$$

d'où $P(\left|\bar{X}_n - m\right| \geq \varepsilon) \underset{n \longrightarrow +\infty}{\longrightarrow} 0$ pour tout $\varepsilon > 0$, ce qui, justement veut dire que $\left(\bar{X}_n\right)$ converge en probabilité vers m.

3-b- On veut établir que $E\left[\left(\bar{X}_n - m\right)^2\right] \underset{n \longrightarrow +\infty}{\longrightarrow} 0$. Pour cela, il suffit de remarquer que $E\left[\left(\bar{X}_n - m\right)^2\right] = E\left[\left(\bar{X}_n - E\left(\bar{X}_n\right)\right)^2\right]$ car $E\left(\bar{X}_n\right) = m$ or $E\left[\left(\bar{X}_n - E\left(\bar{X}_n\right)\right)^2\right] = V(\bar{X}_n)$ donc $E\left[\left(\bar{X}_n - m\right)^2\right] = V(\bar{X}_n) = \frac{\sigma^2}{n} \underset{n \longrightarrow +\infty}{\longrightarrow} 0$. Donc $\left(\bar{X}_n\right)$ converge en moyenne quadratique vers m.

Puisque la convergence en moyenne quadratique implique la convergence en probabilité alors $\left(\bar{X}_n\right)$ converge en probabilité vers m.

En fait, on vient de démontrer dans le cas particulier de la moyenne empirique le résultat suivant qui fournit une condition suffisante de convergence d'un estimateur:

Théorème 1 : Soit $\hat{\theta}_n$ un estimateur de θ. Si $\hat{\theta}_n$ est un estimateur sans biais : $E(\hat{\theta}_n) = \theta$, et que sa variance tend vers 0: $V(\hat{\theta}_n) \underset{n \longrightarrow +\infty}{\rightarrow} 0$ alors $\hat{\theta}_n$ est un estimateur convergent de θ (rappel :lorsqu'on parle d'estimateur convergent sans autre précision, il s'agit de convergence en probabilité)

Ce résultat est lui-même un cas particulier d'un théorème plus général :

Théorème 2 :Soit $\hat{\theta}_n$ un estimateur de θ. Si $\hat{\theta}_n$ est un estimateur asymptotiquement sans biais, c'est-à-dire que $E(\hat{\theta}_n) \underset{n \longrightarrow +\infty}{\rightarrow} \theta$, et que sa variance tend vers 0: $V(\hat{\theta}_n) \underset{n \longrightarrow +\infty}{\rightarrow} 0$ alors $\hat{\theta}_n$ est un estimateur convergent de θ.

3-c- La loi faible des grands nombres reste applicable puisque ses hypothèses sont toujours vérifiées : les X_i sont des variables aléatoires indépendantes admettant la même espérance m et la même variance σ^2. Par contre la loi forte des grands nombres n'est plus applicable puisque l'hypothèse "les X_i suivent la même loi" n'est plus vérifiée.

Par ailleurs, les résultats (et les raisonnements) des questions 3-a et 3-b restent vrais puisqu'ils sont fondés uniquement sur le fait que $E\left(\bar{X}_n\right) = m$ et $V(\bar{X}_n) = \frac{\sigma^2}{n}$, or ces deux résultats restent vrais même sous l'hypothèse plus faible "les X_i sont des variables aléatoires indépendantes admettant la même espérance m et la même variance σ^2".

4-a- On a:

$$P(|U| \leq u) = P(-u \leq U \leq u) = P\left(U \leq u\right) - P(U < -u) = \Phi(u) - \Phi(-u)$$

or la loi normale standard est symétrique par rapport à 0 (au sens où sa densité est symétrique par rapport à 0) donc $\Phi(-u) = 1 - \Phi(u)$. Conclusion :

$$P(|U| \leq u) = 2\Phi(u) - 1$$

4-b- On a les équivalences suivantes :

$$m-\alpha \leq \bar{X}_n \leq m+\alpha \Longleftrightarrow -\alpha \leq \bar{X}_n - m \leq \alpha \Longleftrightarrow -\frac{\alpha}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\alpha}{\frac{\sigma}{\sqrt{n}}} \Longleftrightarrow -\frac{\alpha\sqrt{n}}{\sigma} \leq \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\alpha\sqrt{n}}{\sigma}$$

donc, si on pose $U_n = \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}}$, on a:

$$P\left(m - \alpha \leq \bar{X}_n \leq m + \alpha\right) = P\left(-\frac{\alpha\sqrt{n}}{\sigma} \leq U_n \leq \frac{\alpha\sqrt{n}}{\sigma}\right) = P\left(|U_n| \leq \frac{\alpha\sqrt{n}}{\sigma}\right)$$

or d'après la question 1, $\bar{X}_n \rightsquigarrow N\left(m, \frac{\sigma^2}{n}\right)$ donc $U_n = \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow N\left(0, 1\right)$. L'application du résultat de la question 4-a permet alors d'affirmer que $P\left(|U_n| \leq \frac{\alpha\sqrt{n}}{\sigma}\right) = 2\Phi(\frac{\alpha\sqrt{n}}{\sigma}) - 1$ et par conséquent :

$$P\left(m - \alpha \leq \bar{X}_n \leq m + \alpha\right) = 2\Phi(\frac{\alpha\sqrt{n}}{\sigma}) - 1$$

Lorsque n tend vers l'infini $\Phi(\frac{\alpha\sqrt{n}}{\sigma})$ tend vers $\lim\limits_{u \longrightarrow +\infty} \Phi(u) = 1$ (cette dernière égalité est une propriété générale des fonctions de répartition découlant simplement du fait que la probabilité "totale" est égale à 1). Il en découle que $P\left(m - \alpha \leq \bar{X}_n \leq m + \alpha\right)$ tend vers $2 - 1 = 1$ lorsque n tend vers l'infini. On vient de redémontrer la convergence en probabilité de $\left(\bar{X}_n\right)$ vers m.

4-c- On a les équivalences suivantes :

$$\bar{X}_n - \alpha\frac{\sigma}{\sqrt{n}} \leq m \leq \bar{X}_n + \alpha\frac{\sigma}{\sqrt{n}} \Longleftrightarrow -\alpha\frac{\sigma}{\sqrt{n}} \leq m - \bar{X}_n \leq \alpha\frac{\sigma}{\sqrt{n}} \Longleftrightarrow -\alpha\frac{\sigma}{\sqrt{n}} \leq \bar{X}_n - m \leq \alpha\frac{\sigma}{\sqrt{n}} \Longleftrightarrow -\alpha \leq \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \leq \alpha$$

donc

$$P\left(\bar{X}_n - \alpha\frac{\sigma}{\sqrt{n}} \leq m \leq \bar{X}_n + \alpha\frac{\sigma}{\sqrt{n}}\right) = P\left(-\alpha \leq U_n \leq \alpha\right) = P\left(|U_n| \leq \alpha\right) = 2\Phi(\alpha) - 1$$

On en déduit que pour que $P\left(\bar{X}_n - \alpha\frac{\sigma}{\sqrt{n}} \leq m \leq \bar{X}_n + \alpha\frac{\sigma}{\sqrt{n}}\right) = 1 - \beta$ il faut et il suffit que $2\Phi(\alpha) - 1 = 1 - \beta$ c'est-à-dire que :

$$\Phi(\alpha) = 1 - \frac{\beta}{2}$$

ce qui revient à dire que α est le quantile d'ordre $1 - \frac{\beta}{2}$ de la loi normale standard.

Question-bonus (3 points):

Les X_i sont des variables indépendantes suivant la même loi de Bernouilli $B(1,p)$ donc $n\hat{p}_n = \sum_{i=1}^{n} X_i$ suit la loi binômiale $B(n,p)$. On suppose que l'approximation normale est valable (c'est-à-dire qu'on suppose que $np(1-p) > 15$) donc, on a approximativement : $n\hat{p}_n \rightsquigarrow N(np, np(1-p))$ d'où $\hat{p}_n \rightsquigarrow N\left(p, \frac{p(1-p)}{n}\right)$ et par conséquent, en centrant et en réduisant la variable $\hat{p}_n$, on obtient

$$\frac{\hat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \rightsquigarrow N(0,1)$$

Posons $U_n = \frac{\hat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}}$. Commençons par chercher α tel que $P(|U_n| \leq \alpha) = 0,95$.

On sait que $P(|U_n| \leq \alpha) = 2\Phi(\alpha) - 1$ donc $2\Phi(\alpha) - 1 = 0,95$ soit $\Phi(\alpha) = \frac{1+0.95}{2} = 0,975$ or d'après les données $\Phi(1,96) = 0,975$ donc $\alpha = 1,96$.

Une autre façon (équivalente) pour trouver α est de dire que $P(U_n \leq \alpha) = P(|U_n| \leq \alpha) + \frac{1-P(|U_n|\leq\alpha)}{2}$ car la loi $N(0,1)$ est symétrique par rapport à 0 (retrouvez graphiquement ce résultat) d'où $P(|U_n| \leq \alpha) = 0,95 \Longleftrightarrow P(U_n \leq \alpha) = 0.975$ ce qui d'après l'énoncé nous permet d'affirmer que $\alpha = 1,96$.

Ainsi on a :

$$P\left(\left|\frac{\hat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}}\right| \leq 1,96\right) \simeq 0,95$$

ce qui donne après des manipulations algébriques de base :

$$P\left(-\hat{p}_n - 1,96\sqrt{\frac{p(1-p)}{n}} \leq -p \leq -\hat{p}_n + 1,96\sqrt{\frac{p(1-p)}{n}}\right) \simeq 0,95$$

donc :

$$P\left(\hat{p}_n - 1,96\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p}_n + 1,96\sqrt{\frac{p(1-p)}{n}}\right) \simeq 0,95$$

or les bornes d'un intervalle de confiance ne doivent pas contenir le paramètre de paramètre inconnu (c'est le cas du paramètre p qu'on cherche à estimer) donc on fait l'approximation qui consiste à remplacer p par $\hat{p}_n$ dans les bornes de l'intervalle et on obtient :

$$P\left(\hat{p}_n - 1,96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \leq p \leq \hat{p}_n + 1,96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right) \simeq 0,95$$

ce qui veut dire que $\left[\hat{p}_n - 1,96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \hat{p}_n + 1,96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right]$ est un intervalle (bilatéral) de confiance de niveau (proche de) 95% du paramètre p.