**Statistiques Appliquées. L3**
**Interrogation n°1**

**EXERCICE 1 (2 points) Questions de cours**

**- Est-ce que $V(X-Y) = V(X) - V(Y)$ lorsque X et Y sont indépendantes ?**
Non. Lorsque $X$ et $Y$ sont indépendantes, $V(X-Y) = V(X) + V(Y)$.

**Que vaut $V(X+Y)$ dans le cas où X et Y sont indépendantes et dans le cas où elles ne sont pas indépendantes ?**
Dans le cas général $V(X+Y) = V(X) + V(Y) + 2cov(X, Y)$.
On définit la covariance par $cov(X, Y) = E(XY) - E(X)E(Y)$.

Si $X$ et $Y$ sont indépendantes, alors $cov(X, Y) = 0$ et donc $V(X+Y) = V(X) + V(Y)$.

**- Enoncer le théorème central limite (TCL).**
Si $(X_n)$ est une suite de v.a. indépendantes et de même loi, admettant des moments d'ordre un et deux notés $m = E(X_n)$ et $\sigma^2 = V(X_n)$, alors :

$$\sqrt{n}\frac{\bar{X}_n - m}{\sigma} \rightarrow_{loi} N(0,1)$$

**EXERCICE 2 (2 points) Formule de Bayes**

**La production de téléphones portables est assurée par 2 usines $U_1$ et $U_2$ qui fabriquent respectivement 45% et 55% de la production totale. La probabilité qu'un téléphone soit défectueux sachant qu'il provient de $U_1$ est de 4% et la probabilité qu'un téléphone soit défectueux sachant qu'il provient de $U_2$ est de 3%.**
**Quelle est la probabilité qu'un téléphone choisi au hasard provienne de $U_1$ sachant qu'il est défectueux ?**

On note D l'évènement "le téléphone portable est défectueux".

$P(U_1/D) = \frac{P(D/U_1)P(U_1)}{P(D/U_1)P(U_1)+P(D/U_2)P(U_2)}$

avec $P(D/U_1) = 0.04$, $P(U_1) = 0.45$, $P(D/U_2) = 0.03$ et $P(U_2) = 0.55$. Ainsi :

$$P(U_1/D) = \frac{0,4 * 0,45}{0,4 * 0,45 + 0,3 * 0,55} \approx 0,57$$

**EXERCICE 3 (5 points) Fonction de densité conjointe**

**On considère la densité de probabilité de $(X, Y)$ :**

$$f_{X,Y}(x,y) = \begin{cases} kxy \text{ si } x \in [0,2] \text{ et } y \in [0,5] \\ 0 \text{ sinon} \end{cases}$$

**1. Calculer la constante $k$**

L'intégrale de $f$ sur $R^2$ doit être égale à 1 :
$1 = \int\int kxydxdy = k\int\int xydxdy = k(\int_0^2 xdx)(\int_0^5 ydy) = k[\frac{x^2}{2}]_0^2[\frac{y^2}{2}]_0^5 = 25k$
donc $k = \frac{1}{25}$
donc

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{25}xy \text{ si } x \in [0,2] \text{ et } y \in [0,5] \\ 0 \text{ sinon} \end{cases}$$

**2. Déterminer les densités marginales $f_X(x)$ et $f_Y(y)$ de $X$ et de $Y$. Les v.a.r sont-elles indépendantes ?**

Si $x < 0$ ou $x > 2$, $f_X(x) = 0$
Si $x \in [0,2]$, $f_X(x) = \int_0^5 \frac{1}{25}xydy = \frac{x}{25}\int_0^5 ydy = \frac{x}{2}$

Si $y < 0$ ou $y > 5$, $f_Y(y) = 0$
Si $y \in [0,5]$, $f_Y(y) = \int_0^2 \frac{1}{25}xydx = \frac{y}{25}\int_0^2 xdx = \frac{2y}{25}$

$f_X(x)f_Y(y) = f_{X,Y}(x,y)\ \forall x \in R, \forall y \in R$ donc les variables X et Y sont indépendantes.

**3. Déterminer, lorsqu'elles existent, $f_X(x|Y=y)$ et $f_Y(y|X=x)$ les densités conditionnelles de X sachant $\{Y=y\}$ et de Y sachant $\{X=x\}$.**

Les variables aléatoires $X$ et $Y$ sont indépendantes donc $f_X(x|Y=y) = f_X(x)$ et $f_Y(y|X=x) = f_Y(y)\ \forall x \in R, \forall y \in R$.

**EXERCICE 4 (5 points) Convergence en probabilité**

**Soit $(X_n)$ une suite de variables aléatoires i.i.d suivant une loi de Bernouilli de paramètre $p$, soit $X_1,...X_n \hookrightarrow i.i.d.B(1,p)$. On pose $S_n = X_1 + ... + X_n$.**

**1. Calculer $E(\frac{S_n}{n})$ et $V(\frac{S_n}{n})$.**

$X_1,...X_n \hookrightarrow i.i.d.B(1,p)$ donc pour $1 \le i \le n$, $E(X_i) = p$ et $V(X_i) = p(1-p)$.

$E(\frac{S_n}{n}) = E(\frac{\sum_{i=1}^n X_i}{n}) = \frac{1}{n}.n.E(X_i) = E(X_i) = p$

$V(\frac{S_n}{n}) = V(\frac{\sum_{i=1}^n X_i}{n}) = \frac{1}{n^2}V(\sum_{i=1}^n X_i) = \frac{1}{n^2}\sum_{i=1}^n V(X_i) = \frac{1}{n^2}.n.V(X_i) = \frac{p(1-p)}{n}$

**2. En utilisant l'inégalité de Bienaymé-Tchebychev, montrer que $\frac{S_n}{n}$ converge en probabilité vers $p$.**

On dit que la suite $(\frac{S_n}{n})$ converge en probabilité vers $p$ (noté $\frac{S_n}{n} \rightarrow_p p$) lorsque pour tout $\epsilon > 0$

$$lim_{n\rightarrow+\infty}P(|\frac{S_n}{n} - p| > \epsilon) = 0$$

D'après l'inégalité de Bienaymé-Chebychev, pour tout $\epsilon > 0$ fixé

$$P(|\frac{S_n}{n} - E(\frac{S_n}{n})| \geq \epsilon) \leq \frac{V(\frac{S_n}{n})}{\epsilon^2}$$

$$P(|\frac{S_n}{n} - p| > \epsilon) \leq \frac{p(1-p)}{n\epsilon^2}$$

Donc

$$lim_{n\rightarrow+\infty}P(|\frac{S_n}{n} - p| > \epsilon) \leq lim_{n\rightarrow+\infty}\frac{p(1-p)}{n\epsilon^2}$$

En outre une probabilité est toujours positive, donc

$$0 \leq lim_{n\rightarrow+\infty}P(|\frac{S_n}{n} - p| > \epsilon) \leq lim_{n\rightarrow+\infty}\frac{p(1-p)}{n\epsilon^2}$$

Or $lim_{n\rightarrow+\infty}\frac{p(1-p)}{n\epsilon^2} = 0$
Donc d'après le théorème d'encadrement

$$lim_{n\rightarrow+\infty}P(|\frac{S_n}{n} - p| > \epsilon) = 0$$

Ainsi $\frac{S_n}{n} \rightarrow_p p$.
On a montré que la fréquence empirique de l'évènement $\bar{X}_n = \frac{S_n}{n}$ converge en probabilité vers la fréquence théorique. C'est le théorème de Bernouilli. Cf Lecoutre p. 174

**EXERCICE 5 (7 points) Estimation par intervalle de confiance, intervalle pour une proportion**

**On s'intéresse à la valeur de la probabilité $p$ d'être en faveur d'un candidat C à une élection.**
**On réalise un sondage sur un échantillon de $n = 1000$ personnes et on observe le nombre de personnes parmi les 1000 qui se prononcent en faveur du candidat C.**
**1. Décrire le modèle statistique : quelles sont les variables aléatoires observées $X_1, ...X_n$ ? Quelle est leur loi ?**

Les variables $X_1, ...X_n$ prennent la valeur 1 si l'individu est en faveur du candidat C et 0 sinon. C'est un échantillon de Bernouilli $X_1, ...X_n \hookrightarrow i.i.d.B(1,p)$.

**2. Quel estimateur $F$ proposez-vous pour $p$ ? Quelle est la loi de $1000F$ ?**

L'estimateur est la fréquence observée $F = \frac{1}{n}\sum_{i=1}^{n} X_i$

Ici $F = \frac{1}{1000}\sum_{i=1}^{1000} X_i$.
$nF \hookrightarrow B(n, p)$
$1000F \hookrightarrow B(1000, p)$.

**3. On suppose que la loi de $1000F$ peut être approchée par une loi normale. Donner ses paramètres. Par quelle loi peut-on approcher la loi de $F$ ?**

Tant que $np(1-p) > 15$, on peut approcher la loi de $nF$ par une loi normale :

$$B(n, p) \approx N(np, np(1-p))$$

$$nF \hookrightarrow N(np, np(1-p))$$

$$1000F \hookrightarrow N(1000p, 1000p(1-p))$$

$$F \hookrightarrow N(p, \frac{p(1-p)}{n})$$

$$F \hookrightarrow N(p, \frac{p(1-p)}{1000})$$

**4. Construire un intervalle de confiance bilatéral à 0.95 pour $p$.**

$$\frac{F-p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0, 1)$$

On fait une seconde approximation en remplaçant $p$ par $F$ au dénominateur :

$$\frac{F-p}{\sqrt{\frac{F(1-F)}{n}}} \approx N(0, 1)$$

On commence par chercher la valeur a telle que
$P(-a < \frac{F-p}{\sqrt{\frac{F(1-F)}{n}}} < a) = 0,95$
$\Leftrightarrow \Phi(a) - \Phi(-a) = 0,95$
$\Leftrightarrow 2\Phi(a) - 1 = 0,95$
$\Leftrightarrow \Phi(a) = 0,975$.
La lecture de la table de la loi normale centrée réduite donne $a = 1,96$
Donc

$$P(-1,96 \le \frac{F-p}{\sqrt{\frac{F(1-F)}{n}}} \le 1,96) = 0,95$$

$$\Leftrightarrow P(F - 1,96\sqrt{\frac{F(1-F)}{n}} \le p \le F + 1,96\sqrt{\frac{F(1-F)}{n}}) = 0,95$$

$$F - 1,96\sqrt{\frac{F(1-F)}{n}} \le p \le F + 1,96\sqrt{\frac{F(1-F)}{n}}$$

Avec un échantillon de taille $n = 1000$, nous obtenons l'intervalle de confiance
$F - 0,062\sqrt{F(1-F)} \leq p \leq F + 0,062\sqrt{F(1-F)}$

**5. Application numérique. On constate que dans l'échantillon 510 personnes se prononcent en faveur du candidat C.**
**- Calculer l'estimation $f$ correspondant aux observations.**

$f = \frac{510}{1000} = 0,51.$

**- Calculer l'intervalle de $p$ correspondant aux observations.**

$0,51 - 0,062\sqrt{0,51(1-0,51)} \leq p \leq 0,51 + 0,062\sqrt{0,51(1-0,51)}$
soit $0,479 \leq p \leq 0,541$.

**- L'approximation normale est acceptable pour les calculs numériques lorsque $np(1-p) > 15$. Vérifier que pour la valeur de p la plus défavorable (dans l'intervalle), l'approximation normale est acceptable.**

La valeur la plus défavorable de $p$ est la valeur (dans l'intervalle) telle que $np(1-p)$ soit "le plus petit possible". Ici la valeur la plus défavorable est la plus grande 0.541 et $1000 * 0.541 * (1 - 0.541) = 248.32 > 15$, donc l'approximation normale est acceptable.