Statistiques Appliquées
Correction du Contrôle Continu N°1
TD N°5

Gwenn PARENT

Questions de cours : (2 points)

1. **Donnez la définition de la convergence en probabilité.** *(1 point)*

   [Convergence en probabilité] On dit que (X_n) converge en probabilité vers X ($X_n \xrightarrow{P} X$ ou $p\lim_{n\to+\infty} X_n = X$) si

   $$\forall \varepsilon > 0, \lim_{n\to+\infty} Pr(|X_n - X| > \varepsilon) = 0$$

2. **Rappelez l'énoncé du théorème Central-Limite.** *(1 point)*

   Si (X_n) est une suite de v.a; indépendantes et de même loi, admettant des moments d'ordres un et deux notés $m = E(X_n)$ et $\sigma^2 = V(X_n)$, alors :

   $$\sqrt{n}\,\frac{\overline{X_n} - m}{\sigma} \xrightarrow{Loi} \mathcal{N}(0,1)$$

Exercice 1 : Formule de Bayes (2 points)

Dans une Coupe de France de football, l'équipe du Paris-Saint-Germain, de première division, estime qu'elle a 3 chances sur 4 de gagner si elle rencontre une équipe de D2 et qu'elle a 1 chance sur 3 de gagner si c'est une équipe de première division. Vous entendez rapidement à la radio parler de la défaite du PSG, mais ne savez pas contre qui l'équipe jouait son match. Sachant que la proportion d'équipes de D2 restant engagées dans la Coupe est *p*, calculez la probabilité que le PSG ait rencontré une équipe de D2, sachant que le club parisien a perdu son match.

Posons d'abord les notations :
Soit G = l'évènement "le PSG gagne son match" et A = l'évènement "le PSG rencontre une équipe de D2", l'énoncé nous indique donc que :

$$P(G \setminus A) = \frac{3}{4} \quad \longrightarrow \quad P(\overline{G} \setminus A) = \frac{1}{4}$$
$$P(G \setminus \overline{A}) = \frac{1}{3} \quad \longrightarrow \quad P(\overline{G} \setminus \overline{A}) = \frac{2}{3}$$
$$p(A) = p \quad \longrightarrow \quad p(\overline{A}) = 1 - p$$

On cherche ici $P(A \setminus \overline{G})$, soit la probabilité que le PSG ait rencontré une équipe de D2 sachant qu'il a perdu son match.

Il suffit d'appliquer la Formule de Bayes :

$$P(A \setminus \overline{G}) = \frac{P(\overline{G} \setminus A)P(A)}{P(\overline{G} \setminus A)P(A) + P(\overline{G} \setminus \overline{A})P(\overline{A})}$$

$$P(A \setminus \overline{G}) = \frac{\frac{1}{4}p}{\frac{1}{4}p + \frac{2}{3}(1-p)} = \frac{\frac{1}{4}p}{p(\frac{1}{4} - \frac{2}{3}) + \frac{2}{3}}$$

$$P(A \setminus \overline{G}) = \frac{\frac{1}{4}p}{\frac{2}{3} - \frac{5}{12}p} = \frac{3p}{8-5p}$$

Exercice 2 : Fonction de densité conjointe (7 points)

Soit (X, Y) un couple de variables aléatoires dont la loi est déterminée par la densité conjointe suivante :

$$f_{X,Y}(x,y) = \begin{cases} k\ e^{-(x+y)} & \text{si } (x,y)\ \epsilon\ [0,+\infty] \times [0,x] \\ 0 & \text{sinon} \end{cases}$$

1. **Déterminez la valeur de la constante k.** *(2 points)*
   On sait que k doit vérifier :
   $$\int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} f_{X,Y}(x,y)dy \right) dx = 1$$
   On doit respecter la condition $(x,y)\ \epsilon\ [0,+\infty] \times [0,x]$ pour intégrer, d'où :
   $$k \int_0^{+\infty} e^{-x} \left(\int_0^x e^{-y} dy \right) dx = 1$$
   $$k \int_0^{+\infty} e^{-x} \Big[-e^{-y} \Big]_0^x dx = 1$$
   $$k \int_0^{+\infty} e^{-x}(1 - e^{-x}) dx = 1$$
   $$k \int_0^{+\infty} \left(e^{-x} - e^{-2x} \right) dx = 1$$
   $$k \Big[-e^{-x} + \frac{e^{-2x}}{2} \Big]_0^{+\infty} = 1$$
   $$k(1 - \frac{1}{2}) = \frac{k}{2} = 1 \Leftrightarrow k = 2$$
   La densité s'écrit donc : $f_{X,Y} = 2e^{-(x+y)}$ si $(x,y)\ \epsilon\ [0,+\infty] \times [0,x]$ et $f_{X,Y} = 0$ sinon.

2. **Déterminez les lois marginales de X et de Y. Ces variables sont-elles indépendantes ?** *(2 points)*
   La densité marginale de X est :
   $$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dy \qquad si\ x\ \epsilon\ [0,+\infty],\ 0\ sinon$$

$$f_X(x) = \int_0^x 2e^{-(x+y)}dy \qquad si\ x\ \epsilon\ [0,+\infty],\ 0\ sinon$$

$$f_X(x) = 2e^{-x}\Big[-e^{-y}\Big]_0^x \qquad si\ x\ \epsilon\ [0,+\infty],\ 0\ sinon$$

$$f_X(x) = 2e^{-x}(1-e^{-x}) \qquad si\ x\ \epsilon\ [0,+\infty],\ 0\ sinon$$

De même, la densité marginale de Y est :

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dx \qquad si\ y\ \epsilon\ [0,x],\ 0\ sinon$$

$$f_Y(y) = \int_0^{+\infty} 2e^{-(x+y)}dx \qquad si\ y\ \epsilon\ [0,x],\ 0\ sinon$$

$$f_Y(y) = 2e^{-y}\int_0^{+\infty} e^{-x}dx \qquad si\ y\ \epsilon\ [0,x],\ 0\ sinon$$

$$f_Y(y) = 2e^{-y}\Big[-e^{-x}\Big]_0^{+\infty} \qquad si\ y\ \epsilon\ [0,x],\ 0\ sinon$$

$$f_Y(y) = 2e^{-y}(1) \qquad si\ y\ \epsilon\ [0,x],\ 0\ sinon$$

soit : $f_Y(y) = 2e^{-y}$ si $y\ \epsilon\ [0,x]$, 0 sinon.
On remarque que $f_{X,Y}(x,y) \neq f_X(x) \times f_Y(y)$ (en effet : $2e^{-(x+y)} \neq 2e^{-x}(1-e^{-x}) \times 2e^{-y} = 4e^{-(x+y)}(1-e^{-x})$), ce qui prouve que X et Y ne sont pas indépendants. Cela était déjà suggéré par le fait que $0 < y < x$ (y est contraint par x, donc les variables aléatoires ne sont pas indépendantes)

3. **Déterminez les densités conditionnelles de $X \setminus Y = y$ et de $Y \setminus X = x$.** *(1 point)*
La densité conditionnelle de $X|Y = y$ s'écrit :

$$f_{X|Y=y}(x,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{2e^{-(x+y)}}{2e^{-y}} = e^{-x}\ si\ (x,y)\ \epsilon\ [0,+\infty] \times [0,x]$$

La densité conditionnelle de $Y|X = x$ s'écrit :

$$f_{Y|X=x}(x,y) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{2e^{-(x+y)}}{2e^{-x}(1-e^{-x})} = \frac{e^{-y}}{(1-e^{-x})}\ si\ (x,y)\ \epsilon\ [0,+\infty] \times [0,x]$$

4. **Déterminez $E(X \setminus Y = y)$.** *(2 points)*

$$E(X|Y=y) = \int_{-\infty}^{+\infty} x.f_{X/Y}(x,y)dx = \int_0^{+\infty} x.e^{-x}dx$$

par intégration par partie ($u = x => u' = 1$ et $v' = e^{-x} => v = -e^{-x}$), on a :

$$E(X|Y=y) = \Big[-xe^{-x}\Big]_0^{+\infty} + \int_0^{+\infty} e^{-x}dx$$

$$E(X|Y=y) = 0 + \Big[-e^{-x}\Big]_0^{+\infty} = 1$$

Exercice 3 : Intervalles de confiance (6,5 points)

Nous pouvons considérer que la note finale sur 20 obtenue au contrôle continu de Statistiques est, pour chaque étudiant, une variable aléatoire Y, de loi Normale. Les notes obtenues l'année dernière par 5 groupes de TD (comprenant un total de 127 étudiants) sont supposées être des variables (Y_1, ..., Y_{127}) indépendantes et identiquement distribuées selon une loi Normale, d'espérance m et de variance $\sigma^2 = 13.08 = (3.617)^2$.

1. **Quel est le meilleur estimateur de m ? Quelle est la loi de cet estimateur ?** *(1,5 points)*

   Si $Y_1, ..., Y_N \approx i.i.d.\ N\left(m, \sigma^2\right)$, alors le meilleur estimateur de m est la moyenne empirique : $\overline{Y}_N = \frac{\sum_{i=1}^N Y_i}{N}$
   La loi de cet estimateur est : $\overline{Y}_N = \frac{\sum_{i=1}^N Y_i}{N} \approx N\left(m; \frac{\sigma^2}{N}\right)$

   En effet :

$$E(\overline{Y}_N) = E\left(\frac{\sum_{i=1}^N Y_i}{N}\right) = \frac{1}{N}\sum_{i=1}^N E(Y_i) = \frac{1}{N}(Nm) = m$$

$$V(\overline{Y}_N) = V\left(\frac{\sum_{i=1}^N Y_i}{N}\right) = \left(\frac{1}{N}\right)^2 \sum_{i=1}^N V(Y_i) = \frac{1}{N^2}\ (N\sigma^2) = \frac{\sigma^2}{N}$$

2. **Construire un intervalle bilatéral I de confiance à 90% pour m.** *(3 points)*

   $\overline{Y}_N = \frac{\sum_{i=1}^N Y_i}{N} \approx N\left(m; \frac{\sigma^2}{N}\right)$ et donc la fonction pivotale U_N associée est :

$$U_N = \frac{\overline{Y}_N - m}{\sqrt{\frac{\sigma^2}{N}}} \approx N(0;1) \quad \text{(même pour les petites valeurs de N)}$$

   Un intervalle bilatéral de confiance à 90% correspond à une probabilité d'erreur de 5% de chaque côté. (n'hésitez pas à refaire le dessin pour visualiser quelle valeur rechercher dans la table de la loi normale).

   Lecture dans la table de la loi $N(0;1)$: $P(U_N \leq 1,645) = 0.95$ est équivalent à $P(|U_N| \leq 1,645) = 0.90$
   On en déduit que : $P(-1,645 \leq U_N \leq 1,645) = 0,90$
   Résolvons les inégalités en m , afin d'obtenir l'intervalle bilatéral de confiance 90% pour m :

$$P\left(-1,645 \leq \frac{\overline{Y}_N - m}{\sqrt{\frac{\sigma^2}{N}}} \leq 1,645\right) = 0,90$$

$$P\left(-1,645\sqrt{\frac{\sigma^2}{N}} \leq \overline{Y}_N - m \leq 1,645\sqrt{\frac{\sigma^2}{N}}\right) = 0,90$$

$$P\left(\overline{Y}_N - 1,645\sqrt{\frac{\sigma^2}{N}} \leq m \leq \overline{Y}_N + 1,645\sqrt{\frac{\sigma^2}{N}}\right) = 0,90$$

3. **Application numérique : calculer l'intervalle obtenu précédemment sachant que la valeur moyenne observée sur les 5 groupes de TD de l'année dernière est égale à 10.14 :** *(0,5 point)*

$$\overline{Y} = \frac{1}{127}\sum_{i=1}^{127} Y_i = 10.14$$

Application numérique : pour $N = 127$, nous obtenons l'intervalle bilatéral de confiance à 90% suivant :

$$P\left\{\overline{Y}_{127} - 0.528 \leq m \leq \overline{X}_{127} + 0.528\right\} = 0,90$$

Soit :

$$P\left\{9.612 \leq m \leq 10.668\right\} = 0,90$$

4. **Si vous souhaitez prévoir la note qu'obtiendra un élève de Statistiques cette année (étudiant dans la même université que celle des 5 groupes de TD étudiés précédemment), en faisant l'hypothèse que sa note notée Y_{2007} est indépendante des 127 variables déja observées et suit la même loi qu'elles. Indiquez comment vous procéderiez pour calculer un intervalle de prévision pour Y_{2007} (Indiquez les étapes à suivre, et sur quelle variable vous vous baseriez, mais il n'est pas demandé de trouver l'intervalle de prévision).** *(1,5 points)*

Je vous redonne toute la correction (avec un pourcentage de confiance de 95%), mais je vous demandais uniquement d'expliquer le procédé.

On suppose que $Y_{2007} \approx N(m, \sigma^2)$ est indépendante des 127 notes déjà observées dans la même université. La meilleure prévision ponctuelle de Y_{2007} serait m si on en connaissait la valeur exacte. Nous considérons donc la meilleure estimation de m, à savoir $\overline{Y_N}$, obtenu car $\overline{Y_N}$ minimise l'erreur quadratique moyenne (MSE).

Nous nous basons donc sur l'erreur de prévision $Y_{2007} - \overline{Y_N}$ qui suit une loi $N\left(0; \sigma^2(1 + \frac{1}{N})\right)$
car $E(Y_{2007} - \overline{Y_N}) = E(Y_{2007}) - E(\overline{Y_N}) = m - m = 0$
et $V(Y_{2007} - \overline{Y_N}) = V(Y_{2007}) + V(\overline{Y_N}) - 2cov(Y_{2007}; \overline{Y_N}) = \sigma^2 + \frac{\sigma^2}{N} = \sigma^2(1 + \frac{1}{N})$ car Y_{2007} et $\overline{Y_N}$ sont indépendantes.

Donc la fonction pivotale est la suivante :

$$U = \frac{\left(Y_{2007} - \overline{Y_N}\right) - 0}{\sqrt{\sigma^2(1 + \frac{1}{N})}} \approx N(0;1) \quad \text{(même pour les petites valeurs de N)}$$

Un intervalle bilatéral de confiance à 95% correspond à une probabilité d'erreur de 2,5% de chaque côté. Lecture dans la table de la loi $N(0;1)$ pour un intervalle bilatéral de confiance à 95% :
$P(U \leq 1.96) = 0.975$, qui est équivalent à $P(|U| \leq 1.96) = 0.95$:

On en déduit que : $P(-1.96 \leq U \leq 1.96) = 0,95$
Résolvons les inégalités en Y_{2007} , afin d'obtenir l'intervalle bilatéral de confiance 95% pour Y_{2007} :

$$P\left(-1.96 \leq \frac{\left(Y_{2007} - \overline{Y_N}\right)}{\sqrt{\sigma^2\left(1 + \frac{1}{N}\right)}} \leq 1.96\right) = 0,95$$

$$P\left(-1.96\sqrt{\sigma^2\left(1 + \frac{1}{N}\right)} \leq Y_{2007} - \overline{Y_N} \leq 1.96\sqrt{\sigma^2\left(1 + \frac{1}{N}\right)}\right) = 0,95$$

$$P\left(\overline{Y_N} - 1.96\sqrt{\sigma^2\left(1 + \frac{1}{N}\right)} \leq Y_{2007} \leq \overline{Y_N} + 1.96\sqrt{\sigma^2\left(1 + \frac{1}{N}\right)}\right) = 0,95$$

Application numérique : l'intervalle de prévision à 95% pour la note de notre élève de Statistiques cette année est :

$$P\left(\overline{Y_N} - 7.116 \leq Y_{2007} \leq \overline{Y_N} + 7.116\right) = 0,95$$
$$3.024 \leq Y_{2007} \leq 17.256$$

Exercice 4 : Convergence en probabilité et inégalité de Tchebychev (5 points)

Soit (X_1, X_2, ... X_{n+1}) une suite de variables aléatoires de Bernouilli de même paramètre p, deux à deux indépendantes. On introduit pour tout entier naturel non nul $n \varepsilon N^*$, les variables aléatoires :

$$Y_n = \frac{X_n - X_{n+2}}{2} \qquad et \qquad T_n = \frac{Y_1+Y_2+...+Y_n}{n}$$

1. **Déterminez la loi de Y_n : Quelles valeurs peut prendre Y_n ? Donnez la distribution de probabilité de Y_n, et calculez son espérance et sa variance.** *(2 points)*
   Les X_i sont des variables aléatoires de Bernouilli $B(p)$, donc nous avons : $P(X_i = 1) = p$ et $P(X_i = 0) = 1 - p$. Donc Y_n peut prendre les valeurs 0, $\frac{1}{2}$ et $-\frac{1}{2}$, avec les probabilités suivantes :
   $P(Y_n = 0) = P((X_n = 0)\&(X_{n+1} = 0)) + P((X_n = 1)\&(X_{n+1} = 1)) = (1-p)^2 + p^2 = 1 - 2p + 2p^2$
   $P(Y_n = \frac{1}{2}) = P((Y_n = -\frac{1}{2}) = P((X_n = 0)\&(X_{n+1} = 1)) = P((X_n = 1)\&(X_{n+1} = 0))$
   $P(Y_n = \frac{1}{2}) = P((Y_n = -\frac{1}{2}) = \frac{1}{2}(1 - P(Y_n = 0)) = \frac{1}{2}(1 - (1 - 2p + 2p^2)) = p(1-p)$

   $E(Y_n) = E(\frac{X_n - X_{n+2}}{2}) = \frac{1}{2}E(X_n - X_{n+2}) = \frac{1}{2}(p - p) = 0$

   $V(Y_n) = V(\frac{X_n - X_{n+2}}{2}) = (\frac{1}{2})^2 V(X_n - X_{n+2}) = \frac{1}{4}[V(X_n) + V(X_{n+2}) - 2cov(X_n, X_{n+2})]$
   Or les variables X_n et X_{n+2} sont indépendantes donc $cov(X_n, X_{n+2}) = 0$, donc :

   $V(Y_n) = \frac{1}{4}[p(1-p) + p(1-p)] = \frac{p(1-p)}{2}$

2. **Calculez l'espérance de T_n.** *(1 point)*
   $E(T_n) = E(\frac{Y_1+Y_2+...+Y_n}{n}) = \frac{1}{n}E(\sum_{i=1}^{n} Y_i) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}(0) = 0$

3. **On vous donne la variance de T_n : $V(T_n) = \frac{p(1-p)}{n^2}$. En déduire que $T_n \xrightarrow{P} 0$. (T_n converge en probabilité vers 0).** *(2 points)*
   D'après l'inégalité de Bienaymé-Chebychev :

   $$\forall\ k\ >\ 0\ , \quad P(|\ T_n - E(T_n)\ | \geq k) \leq \frac{V(T_n)}{k^2}$$

   Donc :

   $$\forall\ k\ >\ 0\ , \quad P(|\ T_n - 0)\ | \geq k) \leq \frac{\frac{p(1-p)}{n^2}}{k^2}$$

   Soit :

   $$\forall\ k\ >\ 0\ , \quad P(|\ T_n - 0)\ | \geq k) \leq \frac{p(1-p)}{n^2 k^2}$$

   Or $\lim_{n \longrightarrow +\infty} \frac{p(1-p)}{n^2 k^2} = 0$

   Donc d'après le théorème de l'encadrement :

   $$\lim_{n \longrightarrow +\infty} P(|\ T_n - 0)\ | \geq k) = 0$$

   Donc T_n converge en probabilité vers 0 : $T_n \xrightarrow{P} 0$.

Bonus : *Redémontrez que* $V(T_n) = \frac{p(1-p)}{n^2}$ *(+ 1,5 points bonus)*

Faites très attention dans le calcul des variances... ici les Y_i ne sont pas indépendantes entre elles. Il faut donc tenir compte des covariances $cov(Y_i; Y_{i+2})$ ou repasser par les X_i qui elles sont indépendantes.

$V(T_n) = V(\frac{Y_1+Y_2+...+Y_n}{n})$

$V(T_n) = (\frac{1}{n})^2 V(\frac{X_1-X_3}{2} + \frac{X_2-X_4}{2} + \frac{X_3-X_5}{2} + \frac{X_4-X_6}{2} + ... + \frac{X_{n-2}+X_n}{2} + \frac{X_{n-1}+X_{n+1}}{2} + \frac{X_n+X_{n+2}}{2})$

Les X_i vont s'éliminer deux à deux, sauf deux termes au début et deux à la fin :

$V(T_n) = \frac{1}{n^2}(\frac{1}{2})^2 V(\sum_{i=1}^{n}(X_i - X_{i+2}))$

$V(T_n) = (\frac{1}{2n})^2 V(\sum_{i=1}^{n} X_i - \sum_{i=3}^{n+2} X_i))$

$V(T_n) = (\frac{1}{2n})^2 V(X_1 + X_2 - X_{n+1} - X_{n+2})$

Or les variables X_1, X_2, X_{n+1} et X_{n+2} sont indépendantes deux à deux, donc toutes les covariances deux à deux sont égales à 0. Donc :

$V(T_n) = (\frac{1}{2n})^2 \Big(V(X_1) + V(X_2) + V(X_{n+1}) + V(X_{n+2}) \Big)$

$V(T_n) = \frac{1}{4n^2} 4p(1-p) = \frac{p(1-p)}{n^2}$