Statistiques Appliquées
Correction du Contrôle Continu N°1
TD N°3 et 4

Gwenn PARENT

1. **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 :

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

2. **Qu-est ce qu'un estimateur sans biais ?** *(1 point)*

   Le biais de l'estimateur $\widehat{m}$ d'un paramètre m est : $Biais \;=\; E(\widehat{m}) \;-\; m$. On dit d'un estimateur qu'il est non biaisé (ou sans biais), lorsque son biais est égal à 0.

Exercice 1 : Formule de Bayes (2 points)

Un étudiant doit répondre à un questionnaire à choix multiple où cinq réponses sont proposées à une question, une seule étant correcte. Quand l'évènement $A =$ "l'étudiant a bien travaillé pendant l'année" est réalisé, la réponse est fournie avec exactitude. Dans le cas contraire, l'étudiant répond au hasard. Si l'évènement $B =$ "il a fourni la réponse correcte" est réalisé, calculez la probabilité $P(A \setminus B)$ en fonction de $p = P(A)$.

D'après l'énoncé, on a $P(B \setminus A) = 1$, $P(B \setminus \overline{A}) = 1/5$ et $P(A) = p$.
Il suffit d'appliquer la Formule de Bayes :

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

$$P(A \setminus B) = \frac{p}{p + (1-p)/5} = \frac{5p}{4p+1}$$

On peut noter que $P(A \setminus B) \geq P(A) = p$, avec bien sûr $P(A \setminus B) = 0$ si $p = 0$ et $P(A \setminus B) = 1$ si $p = 1$. Dans tous les autres cas, $P(A \setminus B) > P(A)$.

Exercice 2 : Fonction de densité conjointe (4,5 points)

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

$$f_{X,Y}(x,y) = \begin{cases} k(x+y) & \text{si } (x,y) \in [0,1] \times [0,x] \\ 0 & \text{sinon} \end{cases}$$

1. **Déterminez la valeur de la constante k.** *(2 points)*

   On sait que k doit vérifier :
   $$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)dxdy = 1$$
   On doit respecter la condition $(x,y)\ \epsilon\ [0,1]\times[0,x]$ pour intégrer, d'où :
   $$k\int_0^1\left(\int_0^x (x+y)dy\right)dx = 1$$
   $$\iff \quad k\int_0^1\left[\frac{(x+y)^2}{2}\right]_0^x dx = 1 \quad \iff \quad k\int_0^1\left(\frac{(2x)^2}{2}-\frac{x^2}{2}\right)dx = 1$$
   $$\iff \quad \frac{3}{2}\,k\int_0^1 x^2\ dx = 1 \quad \iff \quad \frac{3}{2}\,k\left[\frac{x^3}{3}\right]_0^1 \quad \iff \quad \frac{3}{2}\,k\,\frac{1}{3}=\frac{k}{2}=1$$
   Donc $k=2$, et la densité s'écrit : $f_{X,Y}(x,y)=2(x+y)$ si $(x,y)\ \epsilon\ [0,1]\times[0,x]$ et $f_{X,Y}(x,y)=0$ sinon.

2. **Déterminez les lois marginales de X et de Y. Ces variables sont-elles indépendantes ?** *(1,5 points)*

   Pour $k=2$, la densité marginale de x est :
   $$f_X(x)=\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\ dy \quad si \quad x\ \epsilon\ [0,1],\ 0\ sinon$$
   $$f_X(x)=\int_0^x 2(x+y)\ dy \quad si \quad x\ \epsilon\ [0,1],\ 0\ sinon$$
   $$f_X(x)=\left[(x+y)^2\right]_0^x = (2x)^2-x^2 \quad si \quad x\ \epsilon\ [0,1],\ 0\ sinon$$
   $$f_X(x)=3x^2 \quad si \quad x\ \epsilon\ [0,1],\ 0\ sinon$$
   De même,
   $$f_Y(y)=\int_{-\infty}^{+\infty} f_{X,Y}(x,y)dx \quad si \quad y\ \epsilon\ [0,x],\ 0\ sinon$$
   $$f_Y(y)=\int_0^1 2(x+y)\ dx \quad si \quad y\ \epsilon\ [0,x],\ 0\ sinon$$
   $$f_Y(y)=\left[(x+y)^2\right]_0^1 = (y+1)^2-y^2 = y^2+2y+1-y^2 \quad si \quad y\ \epsilon\ [0,x],\ 0\ sinon$$
   $$f_Y(y)=2y+1 \quad si \quad y\ \epsilon\ [0,x],\ 0\ sinon$$
   On remarque que $f_{X,Y}(x,y)\neq f_X(x)\times f_Y(y)$ (en effet : $2(x+y)\neq 3x^2\times(2y+1)$), ce qui prouve que X et Y ne sont pas indépendantes. Cela nous était déjà suggéré car $0<y<x$ (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 $Y \setminus X = x$.**

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

$$f_{X|Y=y}(x,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{2(x+y)}{2y+1} \quad si\ (x,y) \,\epsilon\, [0,1]\times[0,x],\ et\ 0\ sinon$$

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

$$f_{Y|X=x}(x,y) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{2(x+y)}{3x^2} \quad si\ (x,y) \,\epsilon\, [0,1]\times[0,x],\ et\ 0\ sinon$$

Exercice 3 : Intervalle de confiance (8 points)

Une entreprise de télécommunications cherche à développer un nouveau produit, et charge ses télé-opérateurs de contacter une base de 4000 clients pour leur proposer ce nouvel abonnement. Ils obtiennent 720 retours positifs de la part de leurs clients. Nous nous intéressons à la probabilité p de réponse favorable pour chaque appel et nous pouvons supposer que les comportements des individus sont indépendants les uns des autres.

1. **Décrire, avec précision le modèle statistique correspondant à ces observations : quelle est la variable aléatoire observée, quel est le nombre d'observations, quelle est la loi de ces observations ?**

   La variable aléatoire observée X_i est une variable de Bernouilli valant 1 si le client i répond favorablement à l'appel du télé-opérateur, 0 sinon. Nous disposons de 4000 observations indépendantes et identiquement distribuées (tirage sans remise mais dans une très large population, ce qui est équivalent à un tirage avec remise, et les comportements sont indépendants). On peut donc considérer que l'on a un échantillon de taille 4000 d'une Bernouilli (p).

2. **Quel estimateur proposez-vous pour p ? Quelle est l'estimation correspondant aux observations faites ?**

   L'estimateur optimal pour p est la fréquence observée :

$$\hat{p} = \frac{1}{4000}\sum_{i=1}^{4000} X_i = F$$

   Les observations faites conduisent à $f = \frac{720}{4000} = 0,18$.

3. **Quelle est la loi de cette estimateur ? A quelle condition peut-on utiliser l'approximation normale ? Vérifiez cette condition, et donnez l'estimateur qui sera utilisé.**

   Loi de cet estimateur :

$$\sum_{i=1}^{4000} X_i \;=\; 4000F \longrightarrow Binomiale\ (4000,\ p)$$

   Rappels sur la loi Binomiale $\mathcal{B}(n,p)$:

- *On réalise n épreuves successives indépendantes où on observe à chaque fois la réalisation ou non d'un certain événement A et on note X le nombre de réalisation de A.*

$$P(X = k) = C_n^k p^k (1-p)^{n-k}$$

- *Si $X_i \longrightarrow B(1,p)$, alors $\sum_i^n X_i \longrightarrow \mathcal{B}(n,p)$.*
- *Si $n > 30$ et $np < 15$, alors la loi $\mathcal{B}(n,p)$ converge en loi vers la loi de Poisson : $\mathcal{P}(np)$.*
- *Si $np\ (1-p) > 15$, la loi $\mathcal{B}(np)$ converge en loi vers la loi normale : $\mathcal{N}(np, np(1-p))$.*

Nous pouvons utiliser l'approximation normale pour les calculs numériques, car $np(1-p) = 4000 \times 0,18 \times (1-0,18) = 590,4 > 15$.

$$\text{Donc : } nF \longrightarrow N\left(np,\ np(1-p)\right) \text{ ou encore } F \longrightarrow N\left(p,\ \tfrac{p(1-p)}{n}\right)$$

Redémontrons rapidement cela :

Si : $nF \longrightarrow N\left(np,\ np(1-p)\right)$ alors $E(nF) = np$ et $V(nF) = np(1-p)$

Calculons $E(F)$ et $V(F)$:

$$E(F) = E\left(\frac{1}{n}(nF)\right) = \frac{1}{n}E(nF) = \frac{1}{n}(np) = p$$

$$V(F) = V\left(\frac{1}{n}(nF)\right) = \left(\frac{1}{n}\right)^2 V(nF) = \frac{1}{n^2}\ (np(1-p)) = \frac{p(1-p)}{n}$$

Nous avons donc bien : $F \longrightarrow N\left(p,\ \frac{p(1-p)}{n}\right)$

4. **Construisez un intervalle bilatéral de confiance à 90% pour p, et calculez l'intervalle ici observé.**

Intervalle bilatéral de confiance proche de 90 % pour p :

$$F \longrightarrow N\left(p, \frac{p(1-p)}{n}\right)$$

Soit U la fonction pivotale associée F : $U = \frac{F-p}{\sqrt{\frac{p(1-p)}{n}}} \approx \mathcal{N}(0,1)$ (donné par le théorème Central-Limite).

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

On lit dans la table de la loi Normale centrée réduite $P(|U| < 1,645) = 2F(1,645) - 1 = 0,90$
Attention, il s'agit d'un intervalle bilatéral de confiance, il faut donc rechercher la valeur 0,95 dans la table de la loi Normale Centrée Réduite : $P(U < 1,645) = 0,95$ donc $P(|U| < 1,645) = 0,90$ (refaites le graphique si nécessaire) d'où :

$$p\left(\left|\frac{F-p}{\sqrt{\frac{p(1-p)}{4000}}}\right| < 1,645\right) = 0,90$$

Pour approcher l'inégalité nous remplaçons p au dénominateur par son estimation F (utilisation de la variance empirique $\frac{F(1-F)}{4000}$) et nous obtenons :

$$-1,645 < \frac{F-p}{\sqrt{\frac{F(1-F)}{4000}}} < 1,645$$

$$-1,645\sqrt{\frac{F(1-F)}{4000}} < F-p < 1,645\sqrt{\frac{F(1-F)}{4000}}$$

$$p\left(F-1,645\sqrt{\frac{F(1-F)}{4000}} < p < F+1,645\sqrt{\frac{F(1-F)}{4000}}\right) \simeq 0,90$$

$$f-1,645\sqrt{\frac{f(1-f)}{4000}} < p < f+1,645\sqrt{\frac{f(1-f)}{4000}}$$

Réalisation observée : Intervalle de confiance bilatéral pour p à 90% : $0,17 < p < 0,35$

Exercice 4 : Convergence en probabilité et inégalité de Tchebychev (6 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 :

$$Y_n = \frac{X_n+X_{n+1}}{2} \qquad et \qquad T_n = \frac{Y_1+Y_2+...+Y_n}{n}$$

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 1, avec les probabilités suivantes :
   $P(Y_n = 0) = P((X_n = 0)\&(X_{n+1} = 0)) = (1-p)\times(1-p) = (1-p)^2$
   $P(Y_n = 1) = P((X_n = 1)\&(X_{n+1} = 1)) = p\times p = p^2$
   $P(Y_n = \frac{1}{2}) = 1 - P(Y_n = 1) - P(Y_n = 0) = 1 - p^2 - (1-p)^2 = 1 - p^2 - (1-2p+p^2) = 2p(1-p)$

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

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

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

2. **Calculez l'espérance et la variance de T_n. *(Attention, les Y_i ne sont pas indépendantes)***

   $E(T_n) = E(\frac{Y_1+Y_2+...+Y_n}{n}) = \frac{1}{n}E(\sum_{i=1}^{n} Y_i) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}(np) = p$

   Faites très attention dans le calcul des variances... ici les Y_i et Y_{i+1} ne sont pas indépendantes. Il faut donc tenir compte des covariances $cov(Y_i; Y_{i+1})$.

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

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

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

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

$V(T_n) = \frac{1}{n^2}[V(\frac{X_1}{2}) + V(X_2) + V(X_3) + ... + V(X_n) + V(\frac{X_{n+1}}{2})]$

$V(T_n) = \frac{1}{n^2}[V(\frac{X_1}{2}) + V(\frac{X_{n+1}}{2}) + \sum_{i=2}^{n} V(X_i)]$

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

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

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

3. **En déduire que $T_n \xrightarrow{P} p$. (T_n converge en probabilité vers p)**

   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 :

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

Soit :

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

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

Donc d'après le théorème de l'encadrement :

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

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