Statistiques Appliquées
Correction du Contrôle Continu N°1
TD N°8 et 9

Mathieu VALDENAIRE

Exercice 1 (2,5 points)

Dans un pays donné et à l'occasion d'une élection, on sait que 70% des électeurs résident dans une zone urbaine, et 30% en zone rurale. On a constaté par ailleurs que parmi les habitants des zones urbaines, 60% des électeurs votent démocrate et 40% républicain. Enfin, les démocrates ont recueilli 52,5% des suffrages sur l'ensemble de la population.
Lorsque l'on tire au hasard un électeur démocrate, quelle est la probabilité que celui-ci réside dans une zone urbaine ?

Soit U l'événement "vivre dans une zone urbaine", $\overline{U}$ l'événement "vivre dans une zone rurale", D l'événement "voter démocrate" et $\overline{D}$ l'événement "voter républicain". D'après l'énoncé, on a $P(U) = 0,70$ et donc $P(\overline{U}) = 1 - P(U) = 0,30$. Par ailleurs on sait que $P(D) = 0,525$ et que $P(D \setminus U) = 0,60$
On cherche $P(U \setminus D)$. Par application de la formule de Bayes :

$$P(U \setminus D) = \frac{P(U \cap D)}{P(D)} = \frac{P(D \setminus U)P(U)}{P(D \setminus U)P(U) + P(D \setminus \overline{U})P(\overline{U})}$$

On remarque que $P(D) = P(D \setminus U)P(U) + P(D \setminus \overline{U})P(\overline{U}) = 0,525 = 0,60.0,70 + 0,35.0,30$.
D'où :

$$P(U \setminus D) = \frac{0,60.0,70}{0,525} = 0,80$$

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

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

$$f_{X,Y}(x,y) = \begin{cases} 2\ e^{-(x+y)} & \text{si } (x,y) \in [0,+\infty] \times [0,x] \\ 0 & \text{sinon} \end{cases}$$

Déterminez les lois marginales de X et de Y. Ces variables sont-elles indépendantes ?
La densité marginale de X est :

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

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

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

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

De même, la densité marginale de Y est :

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

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

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

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

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

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)

Exercice 3 : Intervalles de confiance (7 points)

Nous pouvons considérer que la note finale sur 20 obtenue au baccalauréat est, pour chaque élève, une variable aléatoire Y, suivant une loi Normale. Les notes obtenues au baccalauréat en 2003 par 36 élèves de terminale S sont supposées être des variables (Y_1, ..., Y_{36}) indépendantes et identiquement distribuées selon une loi Normale, d'espérance m et de variance $\sigma^2 = 7,29 = (2,7)^2$.

1. **Quel est le meilleur estimateur de m ? Quelle est la loi de cet estimateur ? Expliquer l'intérêt de construire un intervalle de confiance.**

   Le meilleur estimateur de l'espérance m est la moyenne empirique $\overline{Y}_N = \frac{1}{N}\sum_{i=1}^{N} Y_i$

   Cet estimateur suit une loi normale, puisqu'il est une moyenne de variables suivant des lois normales, d'après l'énoncé. Les paramètres de cette loi sont :

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

   et

   $$\begin{aligned} V(\overline{Y}_N) = V(\tfrac{1}{N}\sum_{i=1}^{N} Y_i) &= \tfrac{1}{N^2}V(\sum Y_i) \\ &= \tfrac{1}{N^2}\sum V(Y_i) \text{ car les } Y_i \text{ sont indépendants} \\ &= \tfrac{1}{N^2}NV(Y_i) \\ &= \tfrac{1}{N^2}N\sigma^2 \\ &= \tfrac{\sigma^2}{N} \end{aligned}$$

   Ainsi si $Y_1, ..., Y_N \approx i.i.d.N\left(m, \sigma^2\right)$, alors $\overline{Y}_N = \frac{\sum_{i=1}^{N} Y_i}{N} \approx N\left(m; \frac{\sigma^2}{N}\right)$ et donc

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

L'intérêt du calcul d'un intervalle de confiance est de refléter le degré de certitude que nous apporte l'information issue d'un échantillon sur la valeur du paramètre que l'on cherche à estimer. Ainsi pour l'estimation d'une espérance, le fait de calculer un intervalle de confiance permet d'ajouter à l'information apportée par la moyenne une information sur la dispersion des observations qui aboutit à déterminer des bornes entre lesquels on peut être sûr à $x\%$ que la vraie valeur du paramètre se situera.

2. **Construire un intervalle bilatéral I_1 de confiance à 90% pour m.**

Modèle : $Y_1, ..., Y_N \approx i.i.d.N(m, (2,7)^2)$.
Si $Y_1, ..., Y_N \approx i.i.d.N\left(m, \sigma^2\right)$, alors $\overline{Y}_N = \frac{\sum_{i=1}^{N} Y_i}{N} \approx N\left(m; \frac{\sigma^2}{N}\right)$ et donc

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

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 $N(0;1)$: $P\{U_N \leq 1,645\} = 0.95$
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 :

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

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

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

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

3. **Application numérique : calculer l'intervalle obtenu précédemment sachant que la valeur moyenne observée est égale à 11.1 :**

$$\overline{Y} = \frac{1}{36}\sum_{i=1}^{36} Y_i = 11,1 \tag{1}$$

Application numérique : pour $N = 36$, nous obtenons l'intervalle bilatéral de confiance à 90% suivant :

$$P\left\{\overline{Y}_{36} - 1,645\frac{2,7}{6} \leq m \leq \overline{Y}_{36} + 1,645\frac{2,7}{6}\right\} = 0,90$$

$$P\left\{\overline{Y}_{36} - 0,740 \leq m \leq \overline{Y}_{36} + 0,740\right\} = 0,90$$

Soit :

$$P\{10,36 \leq m \leq 11,84\} = 0,90$$

Interprétation : le fait d'observer, sur un échantillon de 36 individus, une note moyenne de 11,1 permet donc d'affirmer avec 90% de confiance, que l'espérance de cette note (dans l'ensemble de la population) est comprise entre 10,36 et 11,84.

4. **De quoi dépend la taille de cet intervalle de confiance ? Calculer un intervalle bilatéral I_2 de confiance à 95% pour m et comparer les résultats obtenus avec ceux de la question précédente.**

L'expression de l'intervalle de confiance calculé ci-dessus était :

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

La longueur de l'intervalle (c'est à dire la distance entre ses deux bornes, soit $2 * (1,645\sqrt{\frac{\sigma^2}{N}})$ dépend donc :
- du niveau de confiance exigé : ici le seuil 1,645 reflète le fait que l'on exige un niveau de confiance de 90% : plus le niveau de confiance est élevé plus l'intervalle sera grand.
- du nombre d'observations : plus le nombre d'observations est grand, plus le terme $\frac{\sigma^2}{N}$ sera petit, et donc plus l'intervalle sera petit. Un grand nombre d'observations augmente la précision de l'estimation de la moyenne empirique.
- de la dispersion de la variable étudiée (σ), : plus la dispersion autour de la moyenne est élevée, plus l'intervalle de confiance devra être grand pour obtenir un niveau de confiance donné.

Si l'on modifie le niveau de confiance exigé, seul le seuil $u_{\alpha/2}$ (ci-dessus, 1,645) est modifié.
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 $N(0;1)$: $P\{U_N \leq 1,96\} = 0.975$
L'expression du nouvel intervalle de confiance est donc :

$$P\left(\overline{Y}_N - 1,96\sqrt{\frac{\sigma^2}{N}} \leq m \leq \overline{Y}_N + 1,96\sqrt{\frac{\sigma^2}{N}}\right) = 0,90$$

Application numérique :
On a toujours $overline Y = \frac{1}{36}\sum_{i=1}^{36} Y_i = 11,1$ Ainsi pour $N = 36$, nous obtenons l'intervalle bilatéral de confiance à 95% suivant :

$$P\left\{\overline{Y}_{36} - 1,96\frac{2,7}{6} \leq m \leq \overline{Y}_{36} + 1,96\frac{2,7}{6}\right\} = 0,95$$

$$P\left\{\overline{Y}_{36} - 0,882 \leq m \leq \overline{X}_{36} + 0,882\right\} = 0,95$$

Soit :

$$P\{10,218 \leq m \leq 11,982\} = 0,95$$

Le fait d'observer, sur un échantillon de 36 individus, une note moyenne de 11,1 permet donc d'affirmer

avec 95% de confiance, que l'espérance de cette note (dans l'ensemble de la population) est comprise entre 10,218 et 11,982.
La longueur de ce second intervalle, à 95% de confiance ($2*0,882 = 1,764$) est supérieure à celle de l'intervalle précédent à 90% de confiance ($2*0,740 = 1,480$) : si, toutes choses égales par ailleurs, on augmente le niveau de confiance exigé (c'est à dire augmenter la probabilité que la moyenne empirique observée sur un échantillon particulier soit dans l'intervalle), alors on doit augmenter la longueur de l'intervalle.

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 :
   $P(Y_n = 0) = P((X_n = 0)\&(X_{n+1} = 0)) + P((X_n = 1)\&(X_{n+1} = 1)) = (1-p)^2 + p^2$
   $P(Y_n = \frac{1}{2}) = p(1-p)$
   $P(Y_n = -\frac{1}{2}) = p(1-p)$

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

   $V(Y_n) = V(\frac{X_n - X_{n+2}}{2}) = (\frac{1}{2})^2 V(X_n - X_{n+2}) = \frac{1}{4}[V(X_n) + V(X_{n+2}) - 2cov(X_n, X_{n+2})]$
   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 $T_n \underrightarrow{P} 0$. ($T_n$ 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 :

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

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

Donc d'après le théorème de l'encadrement :

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

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