4.7 Application du TCL : intervalles de confiance

On considère une suite $X_1, X_2, \ldots, X_n$ de v.a.r idépendantes, suivant toutes une même loi de Bernoulli $\mathscr{B}(p)$ pour un paramètre inconnu $p \in]0,1[$. On cherche à **estimer** p **à travers un échantillon** $(x_1, \ldots, x_n)$ **de** $X_1, \ldots, X_n$; un tel échantillon est une réalisation de $X_1, \ldots, X_n$, c-à-d

$$(x_1, \ldots, x_n) = (X_1(\omega), \ldots, X_n(\omega))$$

pour un $\omega \in \Omega$.

Soit

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Pour n suffisamment grand (dans la pratique $n \geq 30$ suffit), on peut considérer, par le Théorème 6.2.3 que

$$Z := \frac{\overline{X} - p}{\sqrt{p(1-p)}/\sqrt{n}} = \sqrt{n} \frac{\overline{X} - p}{\sqrt{p(1-p)}} \sim \mathscr{N}(0,1).$$

Soit $\alpha \in]0,1[$ fixé (**"seuil de risque"**). On note z_α le nombre réel ≥ 0 tel que $\Pi(-z_\alpha) = \alpha$, c-à-d $\mathbf{P}(Z \geq z_\alpha) = \alpha$ ou bien $\mathbf{P}(Z \leq z_\alpha) = 1 - \alpha$. La valeur z_α est lue sur la table de la loi normale (voir table dans l'Annexe 8 plus loin). Alors

$$\mathbf{P}(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = \Pi(z_{\alpha/2}) - \Pi(-z_{\alpha/2}) = 2\Pi(z_{\alpha/2}) - 1 = 1 - \alpha.$$

et donc

$$\begin{aligned}
1 - \alpha &= \mathbf{P}(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) \\
&= \mathbf{P}\left(-z_{\alpha/2} \leq \sqrt{n} \frac{\overline{X} - p}{\sqrt{p(1-p)}} \leq z_{\alpha/2}\right) \\
&= \mathbf{P}\left(\overline{X} - z_{\alpha/2} \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq p \leq \overline{X} + z_{\alpha/2} \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right).
\end{aligned}$$

Comme $\sqrt{p(1-p)} \leq 1/2$, on obtient

$$\mathbf{P}\left(\overline{X} - z_{\alpha/2} \frac{1}{2\sqrt{n}} \leq p \leq \overline{X} + z_{\alpha/2} \frac{1}{2\sqrt{n}}\right) \geq 1 - \alpha.$$

Posons $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$, qui est la **moyenne observée** au vu de l'échantillon $x_1, \dots, x_n$. **L'intervalle de confiance**, au risque α, est alors donné par

$$\boxed{\mathrm{I}_\alpha = \left[\overline{x} - z_{\alpha/2}\frac{1}{2\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{1}{2\sqrt{n}}\right].}$$

Ceci signifie que $p \in \mathrm{I}_\alpha$ avec probabilité au moins $1-\alpha$. Pour $\alpha = 0.05$, on a $z_{\alpha/2} = 1.96$ et donc $p \in \left[\overline{x} - 1.96\dfrac{1}{2\sqrt{n}}, \overline{x} + 1.96\dfrac{1}{2\sqrt{n}}\right]$ avec probabilité de 95%

Exemple 4.7.1 ((**Sondage**)) Au soir d'une élection, on veut estimer la proportion (inconnue) p des électeurs du candidat A dans l'ensemble de la population, à travers un sondage portant sur 2500 électeurs. Ici, $n = 2500$ est le nombre d'électeurs sondés (taille de l'échantillon du sondage) et, pour chaque $i \in \{1, \dots, 2500\}$, on a $\mathrm{X}_i = 1$ si l'électeur i dit avoir voté pour A et $\mathrm{X}_i = 0$ sinon.

Supposons que le sondage a révélé 1350 votes pour le candidat A. On a donc

$$\overline{x} = \frac{1300}{2500} = 0.54$$

et l'intervalle de confiance au risque $\alpha = 5\%$ est

$$[0.54 - 1.96 \times (1/2\sqrt{2500}), 0.54 + 1.96 \times (1/2\sqrt{2500})] = [0.52, 0.56];$$

on a donc une fourchette entre 52% et 56% dans la prédiction de votes pour le candidat A, avec probabilité de 95%.

Chapitre 5

Couples de variables aléatoires à densité

Nous avons étudié à la Section 3.3 des couples de variables aléatoires *discrets*. Dans ce chapitre, nous abordons les couples de variables aléatoires *continus*.

5.1 Vecteurs aléatoires à densité

Soit $(\Omega, \mathscr{F}, \mathbf{P})$ un espace probabilisé.

Soit $n \geq 1$. On rappelle qu'un *vecteur aléatoire* de dimension n est une application

$$\mathrm{Z} : \Omega \to \mathbf{R}^n, \omega \mapsto (\mathrm{X}_1(\omega), \ldots, \mathrm{X}_n(\omega)),$$

où $\mathrm{X}_1, \ldots, \mathrm{X}_n$ sont des v.a.r. On note en général Z par $\mathrm{Z} = (\mathrm{X}_1, \ldots, \mathrm{X}_n)$.

Dans le cas $n = 2$, on parlera de *couple aléatoire*, et on le notera souvent $\mathrm{Z} = (\mathrm{X}, \mathrm{Y})$.

On considère la tribu $\mathscr{B}(\mathbf{R}^n) \subset \mathscr{P}(\mathbf{R}^n)$ engendrée par les “rectangles" $\mathrm{I}_1 \times \mathrm{I}_2 \times \cdots \times \mathrm{I}_n$, où les I_i sont des intervalles de $\mathbf{R}$. Chaque partie de $\mathscr{B}(\mathbf{R}^n)$ est dite partie borélienne ou (borélien) de $\mathbf{R}^n$. On dit que $\mathscr{B}(\mathbf{R}^n)$ est la tribu des boréliens de $\mathbf{R}^n$.

Définition 5.1.1 (Loi et fonction de répartition d'un vecteur aléatoire) Soit $\mathrm{Z} = (\mathrm{X}_1, \ldots, \mathrm{X}_n)$ un vecteur aléatoire.
(i) La *loi* de Z (ou la *loi conjointe* de $\mathrm{X}_1, \ldots, \mathrm{X}_n$) est la measure de probabilité $\mathrm{P}_\mathrm{Z} : \mathscr{B}(\mathbf{R}^n) \to [0,1]$ définie par

$$\mathrm{P}_\mathrm{Z}(\mathrm{I}_1 \times \mathrm{I}_2 \times \cdots \times \mathrm{I}_n) = \mathbf{P}(\mathrm{X}_1 \in \mathrm{I}_1, \mathrm{X}_2 \in \mathrm{I}_2, \ldots, \mathrm{X}_n \in \mathrm{I}_n).$$

pour tous intervalles $I_1, I_2, \ldots, I_n$ de $\mathbf{R}$.
(ii) La *fonction de répartition* de Z est la fonction $F_Z : \mathbf{R}^n \to [0,1]$ définie par

$$F_Z(x_1, x_2, \ldots, x_n) = \mathbf{P}(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n).$$

Nous allons introduire maintenant les vecteurs aléatoires continus.

Définition 5.1.2 (Vecteur aléatoire à densité) Soit $Z = (X_1, \ldots, X_n)$ un vecteur aléatoire. On dit que Z est une *vecteur aléatoire à densité* ou *vecteur aléatoire continue* si sa fonction de répartition F_Z s'écrit sous la forme

$$F_Z(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(t_1, \ldots, t_n) dt_1 \cdots dt_n \qquad \text{pour tout} \quad (x_1, \ldots, x_n) \in \mathbf{R}^n$$

pour une fonction $f : \mathbf{R}^n \to \mathbf{R}$, appelée *densité de* Z, avec les propriétés suivantes :

1. $f(x_1, \ldots, x_n) \ge 0$ pour tout $(x_1, \ldots, x_n) \in \mathbf{R}^n$;
2. f est intégrable sur $\mathbf{R}^n$;
3. $\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(t_1, \ldots, t_n) dt_1 \cdots dt_n = 1$.

Exemple 5.1.3 Soit $Z = (X, Y)$ de loi uniforme sur le triangle

$$T = \{(x, y) \in \mathbf{R}^2 \mid x \ge 0, y \ge 0, x + y \le 1\};$$

comme l'aire de T est $1/2$, la densité de Z est $f = 2\mathbf{1}_T$, c-à-d

$$f(x, y) = \begin{cases} 2 & \text{si } (x, y) \in T \\ 0 & \text{si } (x, y) \notin T. \end{cases}$$

Remarque 5.1.4 Soit $Z = (X_1, \ldots, X_n)$ un vecteur aléatoire à densité. On peut calculer ainsi la probabilité d'évènements associés à Z au moyen de la densité f de Z :

$$\forall B \in \mathscr{B}(\mathbf{R}^n) \qquad : \qquad \mathbf{P}(Z \in B) = P_Z(B) = \int_B f(t_1, \ldots, t_n) dt_1 \cdots dt_n.$$

5.2 Lois marginales

Dans toute la suite, on se limitera au cas $\boxed{n=2}$, c-à-d à celui d'un couple aléatoire $Z=(X,Y):\Omega\to\mathbf{R}^2$.

Les *lois marginales* du couple Z sont les lois P_X et P_Y de X et Y.

Proposition 5.2.1 *Soit* $Z=(X,Y)$ *un vecteur aléatoire à densité* f. *Alors les v.a.r.* X *et* Y *sont continues, de densités* f_X *et* f_Y *données par*

$$\boxed{f_X(x)=\int_{-\infty}^{+\infty} f(x,y)dy} \qquad \textit{et} \qquad \boxed{f_y(y)=\int_{-\infty}^{+\infty} f(x,y)dx}$$

pour $x,y\in\mathbf{R}$.

Démonstration Soit $x\in\mathbf{R}$. On a $\{X\le x\}=\{X\le x, Y\in\mathbf{R}\}$. Il s'ensuit que, pour la fonction de répartition F_X de X, on a

$$\begin{aligned}F_X(x)&=\mathbf{P}(X\le x)=\mathbf{P}(X\le x, Y\in]-\infty,+\infty[)\\&=P_Z(]-\infty,x]\times]-\infty,+\infty[)=\int_{-\infty}^{x}\int_{-\infty}^{+\infty} f(t,y)dy\\&=\int_{-\infty}^{x} f_X(t)dy,\end{aligned}$$

où $f_X(t)=\int_{-\infty}^{+\infty} f(t,y)dy$. Ceci montre que X est continue à densité f_X.

De manière similaire, on montre que Y est continue à densité f_Y.■

Exemple 5.2.2 Reprenons l'Exemple 5.1.3. On rappelle que la densité de $Z=(X,Y)$ est $f=2\mathbf{1}_T$, où T est le triangle $\{(x,y)\in\mathbf{R}^2\mid x\ge0, y\ge0, x+y\le1\}$.

La loi marginale f_X de X est donnée par $f_X(x)=0$ pour $x\notin[0,1]$ et

$$f_X(x)=2\int_0^{1-x}\mathbf{1}_T(x,y)dxdy=2(1-x)$$

pour $x\in[0,1]$. De manière similaire, $f_Y(y)=0$ pour $y\notin[0,1]$ et $f_Y(y)=2(1-y)$ pour $y\in[0,1]$.

On rappelle que les v.a.r. X et Y sont indépendantes si et seulement si

$$\forall(x,y)\in\mathbf{R}^2: \qquad \mathbf{P}(X\le x, Y\le y)=\mathbf{P}(X\le x)\mathbf{P}(Y\le y).$$

Proposition 5.2.3 *Soit* $Z = (X, Y)$ *un vecteur aléatoire à densité* f. *Les deux propriétés suivantes sont équivalentes :*
(i) X *et* Y *sont indépendantes;*
(ii) $f(x, y) = f_X(x) f_Y(y)$ *pour tout* $(x, y) \in \mathbf{R}^2$.

Démonstration Par définition, X et Y sont indépendantes si et seulement si $F_Z(x, y) = F_X(x)F_Y(y)$, c-à-d si et seulement si

$$(*) \int_{-\infty}^{x} \int_{-\infty}^{y} f(t,s)\,dt\,ds = \left(\int_{-\infty}^{x} f_X(t)\,dt\right)\left(\int_{-\infty}^{y} f_X(s)\,ds\right)$$

pour tout $(x, y) \in \mathbf{R}^2$.

Supposons que $f(x, y) = f_X(x) f_Y(y)$ pour tout $(x, y) \in \mathbf{R}^2$; alors la relation $(*)$ est satisfaite et X et Y sont donc indépendantes.

Réciproquement, supposons que X et Y sont indépendantes et que donc la relation $(*)$ est satisfaite. Alors, on a

$$\int_{-\infty}^{x} \int_{-\infty}^{y} f(t,s)\,dt\,ds = \int_{-\infty}^{x} \int_{-\infty}^{y} (f_X(t) f_X(s))\,dt\,ds$$

pour tout $(x, y) \in \mathbf{R}^2$. Par identification, on a $f(x, y) = f_X(x) f_Y(y)$ pour tout $(x, y) \in \mathbf{R}^2$. ■

On peut démontrer une formule de transfert analogue à celle pour un couple de v.a.r. discrètes.

Proposition 5.2.4 ***(Formule de transfert)*** *Soit* $Z = (X, Y)$ *un vecteur aléatoire à densité* f. *Soit* $g : \mathbf{R}^2 \to \mathbf{R}$ *une fonction continue (ou plus généralement, mesurable). On suppose que l'intégrale* $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |g(x, y)| f(x, y)\,dx\,dy$ *est finie. Alors la v.a.r.* $g(X, Y)$ *possède une espérance et on a*

$$\boxed{\mathbb{E}(g(X,Y)) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y) f(x, y)\,dx\,dy}$$

Démonstration Omise.

5.3 Covariance

Comme pour un couple de v.a.r. discrètes, on peut définir la covariance d'un couple de variables continues.

Définition 5.3.1 (Covariance d'un couple aléatoire à densité) Soit $Z = (X, Y)$ un vecteur aléatoire à densité. On suppose que X et Y possèdent des variances. La *covariance* de Z (ou de X et Y), notée $\mathrm{Cov}(X, Y)$, est le nombre

$$\boxed{\mathrm{Cov}(X,Y) := \mathbb{E}\left((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\right)}$$

Les v.a.r. X et Y sont dites *non corrélées* si $\mathrm{Cov}(X, Y) = 0$.

La covariance a des propriétés similaires à celle pour un couple de v.a.r. discrètes.

Proposition 5.3.2 *Soit* $Z = (X, Y)$ *un vecteur aléatoire à densité* f. *On suppose que* X *et* Y *possèdent une variance.*
(i) ***(Formule de Koenig-Huygens)*** *On a*

$$\boxed{\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).}$$

(ii) On a

$$\boxed{\mathrm{Cov}(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf(x,y)\,dx\,dy - \left(\int_{-\infty}^{+\infty} xf_X(x)\,dx\right)\left(\int_{-\infty}^{+\infty} yf_X(y)\,dy\right).}$$

(iii) Si X *et* Y *sont indépendantes, alors* $\mathrm{Cov}(X, Y) = 0$; *la réciproque est fausse en général.*

Démonstration
(i) On a, en développant et en utilisant la linéarité de l'espérance

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbb{E}\left((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\right) = \mathbb{E}\left(XY - X\mathbb{E}(Y) - \mathbb{E}(X)Y + \mathbb{E}(X)\mathbb{E}(Y)\right)\\ &= \mathbb{E}(XY) - \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

(ii) Par la formule de transfert (Proposition 5.2.4), on a

$$\mathbb{E}(XY) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf(x,y)\,dx\,dy.$$

ceci montre que (ii) est une conséquence de (i) .

(iii) Supposons que X et Y sont indépendantes. Alors, en utilisant la Proposition 5.2.3, on a

$$\begin{aligned}
\mathbb{E}(XY) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf(x,y)dxdy \\
&= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf_X(x)f_Y(y)dxdy \\
&= \left(\int_{-\infty}^{+\infty} xf_X(x)dx\right)\left(\int_{-\infty}^{+\infty} yf_X(y)dy\right) \\
&= \mathbb{E}(X)\mathbb{E}(Y),
\end{aligned}$$

c-à-d $\mathrm{Cov}(X,Y) = 0$.■

Exemple 5.3.3 Reprenons l'Exemple 5.1.3 et 5.2.2 On rappelle que la loi f de $Z = (X,Y)$ est la loi uniforme sur le triangle T ; les lois marginales f_X et f_Y sont données par $f_X(x) = f_Y(x) = 2\mathbf{1}_{[0,1]}(x)(1-x)$. On calcule que

$$\mathbb{E}(X) = \mathbb{E}(Y) = 2\int_0^1 x(1-x)dx = 1/3$$

et que

$$\mathbb{E}(XY) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf(x,y)dxdy = 2\int_0^1\int_0^{1-x} xydxdy = \int_0^1 x(1-x)^2dx = 1/12$$

D'où $\mathrm{Cov}(X,Y) = (1/3)^2 - 1/12 = -1/36$.