Théorème de Cochran et applications en statistiques

1 Théorème de Cochran

On munit $\mathbb{R}^n$ de sa structure euclidienne canonique.

Théorème 1 (Cochran[1]). *Soit $X = {}^t(X_1, \dots, X_n)$ un vecteur gaussien centré réduit. Pour F un sous-espace vectoriel de $\mathbb{R}^n$ de dimension p, on note P_F (resp. $P_{F^\perp}$) la projection orthogonale sur F (resp. $F^\perp$).*
Alors les vecteurs aléatoires $P_F X$ et $P_{F^\perp} X$ sont gaussiens indépendants de lois

$$P_F X \sim \mathcal{N}(0, P_F) \quad \text{et} \quad P_{F^\perp} X \sim \mathcal{N}(0, P_{F^\perp})$$

De plus, les variables aléatoires $\|P_F X\|^2$ et $\|P_{F^\perp} X\|^2$ sont indépendantes de lois

$$\|P_F X\|^2 \sim \chi^2_p \quad \text{et} \quad \|P_{F^\perp} X\|^2 \sim \chi^2_{n-p}$$

Remarque 2. Ce théorème est un analogue "en loi" du théorème de Pythagore. L'identité $\|x\|^2 = \|P_F x\|^2 + \|P_{F^\perp} x\|^2$ (pour $x \in \mathbb{R}^n$) devient en effet dans le contexte du théorème $\|X\|^2 \overset{\mathcal{L}}{=} \|P_F X\|^2 + \|P_{F^\perp} X\|^2$, et on a aussi (surtout !) les lois des 2 termes de la somme.

Démonstration. Le résultat est immédiat si on l'écrit dans une base orthonormée adaptée à la somme directe orthogonale $\mathbb{R}^n = F \oplus F^\perp$: soit $(u_1, \dots, u_p)$ (resp. $(u_{p+1}, \dots, u_n)$) une base orthonormée de F (resp. $F^\perp$), alors $u = (u_1, \dots, u_n)$ est une base orthonormée de $\mathbb{R}^n$. Notons U la matrice (orthogonale, ${}^t U = U^{-1}$) de passage de la base canonique à la base u.
Les projections orthogonales sur F et $F^\perp$ s'expriment très simplement dans la base u :

$$P_F = U I_p \, {}^t U \quad \text{et} \quad P_{F^\perp} = U J_{n-p} \, {}^t U$$

où I_p est la matrice diagonale avec des 1 sur les p premiers coefficients diagonaux et des 0 ensuite, et $J_{n-p} = \mathrm{Id} - I_p$.

On pose $Y = {}^t U X$. C'est encore un vecteur gaussien centré réduit (car il est de matrice de covariance ${}^t U \mathrm{Id} U = \mathrm{Id}$, la loi gaussienne centrée réduite est invariante par rotation), qui correspond aux coordonnées de X dans la base u.
Pour Y, on a immédiatement que $I_p Y = {}^t(Y_1, \dots, Y_p, 0, \dots, 0)$ et $J_{n-p} Y = {}^t(0, \dots, 0, Y_{p+1}, \dots, Y_n)$ sont indépendants, de lois $\mathcal{N}(0, I_p)$ et $\mathcal{N}(0, J_{n-p})$, puis que $\|I_p Y\|^2 = \sum_{i=1}^p Y_i^2 \sim \chi^2_p$ et $\|J_{n-p} Y\|^2 = \sum_{i=p+1}^n Y_i^2 \sim \chi^2_{n-p}$.

On peut alors revenir au vecteur X en remarquant que $P_F X = U I_p Y$ et $P_{F^\perp} X = U J_{n-p} Y$ sont gaussiens centrés indépendants de matrice de covariance respective $U I_p \, {}^t U = P_F$ et $U J_{n-p} \, {}^t U = P_{F^\perp}$, puis, comme une transformation orthogonale préserve la norme, que

$$\|P_F X\|^2 = \|I_p Y\|^2 \sim \chi^2_p \quad \text{et} \quad \|P_{F^\perp} X\|^2 = \|J_{n-p} Y\|^2 \sim \chi^2_{n-p}$$

□

[1] Je triche un peu, c'est une version simplifiée, donc plus compréhensible mais généralement suffisante en pratique, du théorème de Cochran.

2 Statistique des échantillons gaussiens

Soit $(X_1, \cdots, X_n)$ un échantillon de variables aléatoires réelles iid de loi $\mathcal{N}(\mu, \sigma^2)$.
On note

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{la moyenne empirique de l'échantillon}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2 \qquad \text{la variance empirique de l'échantillon}$$

Théorème 3. *Les variables aléatoires $\overline{X}$ et S^2 sont indépendantes, et on connait les lois de*

$$\overline{X} \sim \mathcal{N}\Big(\mu, \frac{\sigma^2}{n}\Big) \quad ; \quad (n-1)\frac{S^2}{\sigma^2} = \sum_{i=1}^{n} (X_i - \overline{X})^2 \sim \chi^2_{n-1} \quad ; \quad \sqrt{n}\frac{\overline{X}-\mu}{S} \sim T_{n-1}$$

Remarque 4. On note T_n la loi de Student à n degrés de liberté, qui est par définition la loi de $\frac{X}{\sqrt{Z/n}}$, avec X et Z indépendantes, X de loi normale centrée réduite, Z de loi du chi-deux à n degrés de liberté.

Démonstration. Soit $Y = {}^t(Y_1, \ldots, Y_n)$ un vecteur gaussien centré réduit. On notera

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad \text{et} \qquad R^2 = \frac{1}{n-1}\sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

On sait que $\overline{Y}$ est de loi normale centrée de variance $\frac{1}{n}$.

Soit $\mathbf{1} = {}^t(1, \ldots, 1) \in \mathbb{R}^n$ et $F = \text{Vect}(\mathbf{1})$. Pour tout $y \in \mathbb{R}^d$, on note $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. On vérifie immédiatement que $P_F(y) = \bar{y}\mathbf{1}$ car $\bar{y}\mathbf{1} \in F$ et

$$\langle y - \bar{y}\mathbf{1}, \mathbf{1}\rangle = \sum_{i=1}^{n} (y_i - \bar{y}) = 0$$

donc $y - \bar{y} \in F^\perp$. On en déduit que $\overline{Y}\mathbf{1} = P_F Y$ et $Y - \overline{Y}\mathbf{1} = P_{F^\perp}(Y)$. On peut alors appliquer le théorème de Cochran, en remarquant que $F^\perp$ est de dimension $n-1$. Ainsi la variable

$$\big\|Y - \overline{Y}\mathbf{1}\big\|^2 = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = (n-1)R^2 \sim \chi^2_{n-1}$$

et est indépendante de $\overline{Y}$. On en déduit immédiatement, par définition de la loi de Student, que

$$\sqrt{n}\frac{\overline{Y}}{R} \sim T_{n-1}$$

On sait que $X = {}^t(X_1, \cdots, X_n) \overset{\mathcal{L}}{=} \mu + \sigma Y$, soit aussi

$$\overline{X} \overset{\mathcal{L}}{=} \mu + \sigma\overline{Y} \sim \mathcal{N}\Big(\mu, \frac{\sigma^2}{2}\Big)$$

$$X - \overline{X}\mathbf{1} \overset{\mathcal{L}}{=} \sigma(Y - \overline{Y}\mathbf{1}) \in F^\perp$$

$$(n-1)\frac{S^2}{\sigma^2} = \frac{1}{\sigma^2}\big\|X - \overline{X}\mathbf{1}\big\|^2 \overset{\mathcal{L}}{=} \big\|Y - \overline{Y}\mathbf{1}\big\|^2 = (n-1)R^2 \sim \chi^2_{n-1}$$

$$\sqrt{n}\frac{\overline{X}-\mu}{S} \overset{\mathcal{L}}{=} \sqrt{n}\frac{\overline{Y}}{R} \sim T_{n-1}$$

et $\overline{X}$ et S^2 sont indépendantes. □

Corollaire 5. *La variable aléatoire $\overline{X}$ (resp. S^2) est un estimateur sans biais et convergent de μ (resp. σ^2). De plus, la connaissance des lois de $(n-1)\frac{S^2}{\sigma^2}$ et $\sqrt{n}\frac{\overline{X}-\mu}{S}$ permet de construire des intervalles de confiance pour ces estimations.*

3 Modèle linéaire gaussien

Soit $(x_1, \dots, x_n)$ des valeurs fixées, et $(Y_1, \dots, Y_n)$ un échantillons de variables aléatoires réelles définies par $Y_i = \alpha + \beta x_i + \sigma E_i$ où $(E_1, \dots, E_n)$ sont des vaiid gaussiennes centrées réduites.
On peut remarquer que c'est une généralisation du modèle étudié section 2, qui correspond exactement au cas $\beta = 0$ (et $\alpha = \mu$). Ici aussi, le calcul sur les vecteurs gaussiens va permettre de construire des estimateurs et des intervalles de confiance (voire des tests) pour les paramètres du modèle α, β et σ^2.
On note

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$B = \frac{\sum_{i=1}^{n}(x_i - \overline{x})Y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad A = \overline{Y} - B\overline{x}$$

$$\forall\, x_0 \in \mathbb{R}\ ,\ Y_0^\star = A + Bx_0 \qquad S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - Y_i^\star)^2$$

Théorème 6. *Les variables aléatoires $\overline{Y}$, B et S^2 sont indépendantes, et on connait les lois de*

$$\overline{Y} \sim \mathcal{N}\Big(\alpha + \beta\overline{x}, \frac{\sigma^2}{n}\Big) \qquad B \sim \mathcal{N}\Big(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\Big)$$

$$A \sim \mathcal{N}\Big(\alpha, \sigma^2\Big(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\Big)\Big) \qquad Y_0^\star \sim \mathcal{N}\Big(\alpha + \beta x_0, \sigma^2\Big(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\Big)\Big)$$

$$(n-2)\frac{S^2}{\sigma^2} \sim \chi^2_{n-2}$$

Démonstration. Les deux variables aléatoires $\overline{Y}$ et B sont obtenues par combinaison linéaire des $(Y_i)_{1\leq i\leq n}$ gaussiennes indépendantes, donc sont gaussiennes de moyennes $\mathbb{E}(\overline{Y}) = \alpha+\beta\overline{x}$ et $\mathbb{E}(B) = \beta$ et de variances et covariance

$$\mathrm{Var}(\overline{Y}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(Y_i) = \frac{\sigma^2}{n}$$

$$\mathrm{Var}(B) = \frac{1}{\big(\sum_{i=1}^{n}(x_i - \overline{x})^2\big)^2}\sum_{i=1}^{n}(x_i - \overline{x})^2\mathrm{Var}(Y_i) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\mathrm{Cov}(\overline{Y}, B) = \frac{1}{n\sum_{i=1}^{n}(x_i - \overline{x})^2}\sum_{i=1}^{n}(x_i - \overline{x}) = 0$$

Comme $\overline{Y}$ et B sont gaussiennes indépendantes, on obtient immédiatement que $A = \overline{Y} - B\overline{x}$ et $Y_0^\star = A + Bx_0 = \overline{Y} + B(x_0 - \overline{x})$ sont gaussiennes, ainsi que leur loi.

Soit $\mathbf{1} = {}^t(1, \dots, 1)$ et $x - \overline{x}\mathbf{1}$ deux vecteurs (orthogonaux) de $\mathbb{R}^d$, et $F = \mathrm{Vect}(\mathbf{1}, x - \overline{x}\mathbf{1})$. Pour tout $e \in \mathbb{R}^d$, on note $\overline{e} = \frac{1}{n}\sum_{i=1}^{n} e_i$, $b(e) = \frac{\sum_{i=1}^{n}(x_i-\overline{x})e_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$ et $e^\star = \overline{e}\mathbf{1} + b(e)(x - \overline{x}\mathbf{1})$. On a alors $e^\star = P_F(e)$; en effet $e^\star \in F$ et, comme $\langle \mathbf{1}, x - \overline{x}\mathbf{1}\rangle = 0$,

$$\langle e - e^\star, \mathbf{1}\rangle = \langle e - \overline{e}\mathbf{1}, \mathbf{1}\rangle = 0 \qquad \text{et} \qquad \langle e - e^\star, x - \overline{x}\mathbf{1}\rangle = \langle e, x - \overline{x}\mathbf{1}\rangle - b(e)\|x - \overline{x}\mathbf{1}\|^2 = 0$$

On peut donc appliquer le théorème de Cochran au vecteur gaussien centré réduit $E = {}^t(E_1, \dots, E_n)$ pour obtenir que la variable aléatoire $\|E - E^\star\|^2$ suit une loi du chi-deux à $n-2$ degrés de liberté et est indépendante de $\overline{E}$ et $b(E)$.
La conclusion est alors immédiate en remarquant que $Y = \alpha + \beta x + \sigma E$, donc $\overline{Y} = \alpha + \beta\overline{x} + \sigma\overline{E}$, $B = \beta + \sigma b(E)$, $Y^\star = \alpha + \beta x + \sigma E^\star$, et par conséquent $(n-2)\frac{S^2}{\sigma^2} = \|E - E^\star\|^2$. □

Comme dans le cas d'un échantillon gaussien, ce résultat permet de construire des intervalles de confiance centrés en A, B et S^2 pour les paramètres α, β et σ^2. Je détaille en corollaire la construction d'un intervalle de confiance pour $\alpha + \beta x_0$, qui est la moyenne de la variable aléatoire Y_0, lorsqu'une valeur x_0 est donnée. Ça donne une région de confiance pour l'estimation de la droite de liaison linéaire, d'équation $y = \alpha + \beta x$, par la droite de régression linéaire, d'équation $y = A + Bx$.

Corollaire 7. *La variable aléatoire*

$$\frac{Y_0^\star - \alpha - \beta x_0}{S\sqrt{\dfrac{1}{n} + \dfrac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}} \sim T_{n-2}$$

Un intervalle de confiance pour $\alpha + \beta x_0$ est donné par

$$\left[Y_0^\star - t_{n-2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}} \; ; \; Y_0^\star - t_{n-2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}} \right]$$

où t_{n-2} est le quantile de niveau souhaité de la loi de Student à $n-2$ degrés de liberté.

Démonstration. On sait que

$$\frac{Y_0^\star - \alpha - \beta x_0}{\sigma\sqrt{\dfrac{1}{n} + \dfrac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}} \sim \mathcal{N}(0,1) \qquad (n-2)\frac{S^2}{\sigma^2} \sim \chi^2_{n-2}$$

et que S^2 est indépendante de B et de $\overline{Y}$, donc de $Y_0^\star$. On peut donc conclure par définition d'une loi de Student. □

On peut aussi utiliser la valeur estimée $Y_0^\star$ pour prévoir la valeur de $Y_0 = \alpha + \beta x_0 + E_0$ lors d'un tirage futur. Un intervalle de prévision sert à encadrer cette valeur. On utilise pour cela le fait que E_0 est un tirage indépendant de $(E_1, \ldots, E_n)$, donc

$$Y_0 - Y_0^\star \sim \mathcal{N}\left(0; \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)\right)$$

soit aussi

Corollaire 8. *La variable aléatoire*

$$\frac{Y_0 - Y_0^\star}{S\sqrt{1 + \dfrac{1}{n} + \dfrac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}} \sim T_{n-2}$$

Un intervalle de prédiction pour Y_0 est donné par

$$\left[Y_0^\star - t_{n-2} S \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}} \; ; \; Y_0^\star - t_{n-2} S \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}} \right]$$

où t_{n-2} est le quantile de niveau souhaité de la loi de Student à $n-2$ degrés de liberté.

Remarque 9. Par exemple, pour t_{n-2} tel que $\mathbb{P}(|T_{n-2}| > t_{n-2}) = 0.05$, l'intervalle de prédiction contiendra environ (car $Y_0^\star$ n'est qu'une estimation de la vraie moyenne $\alpha + \beta x_0$ au vu des observations précédentes) 95% des tirages de variables indépendantes de même loi que Y_0.