Université de Strasbourg
M1 - Magistère
Statistique - études de cas

Ségolen Geffray
geffray@math.unistra.fr
Année 2015/2016

Sujet 13: Tests de corrélation

• **Quelques rappels sur la définition mathématique des tests d'hypothèses**:
Soit X une variable aléatoire à valeurs dans un ensemble $\mathcal{X}$, de loi P_θ où $\theta \in \Theta$. Supposons que l'on veuille effectuer un test statistique d'hypothèses portant sur θ. On formule deux hypothèses contradictoires notées H_0 et H_1 dont on suppose que l'une et seulement l'une est vraie. Mathématiquement, formuler H_0 et H_1 revient à choisir deux sous-ensembles disjoints de Θ notés Θ_0 et de Θ_1 de sorte que l'hypothèse H_0 s'écrit alors $\{\theta_0 \in \Theta_0\}$ tandis que l'hypothèse H_1 s'écrit $\{\theta_1 \in \Theta_1\}$. Soit $(X_1, ..., X_n)$ un échantillon i.i.d. issu d'une variable parente X et soit $(x_1, ..., x_n) \in \mathcal{X}^n$ une réalisation de $(X_1, ..., X_n)$. Construire un test de H_0 contre H_1 revient à construire une région critique $\mathcal{R}$ de telle sorte que l'on rejette H_0 lorsque $(x_1, ..., x_n) \in \mathbb{R}$ et que l'on s'assure que le risque de 1ère espèce est inférieur ou égal à α.

Dans ce cadre, on parle de **fonction de risque de 1ère espèce** définie sur Θ_0 pour

$$\theta \to \alpha(\theta) = \mathbb{P}_\theta((X_1, ..., X_n) \in \mathcal{R})\,.$$

On appelle **taille du test** la probabilité maximale de rejeter H_0 alors que H_0 est vraie:

$$\sup_{\theta \in \Theta_0} \{\mathbb{P}_\theta((X_1, ..., X_n) \in \mathcal{R})\}\,.$$

On dit qu'un test est de **niveau** α si sa taille est égale à α ie si

$$\sup_{\theta \in \Theta_0} \{\mathbb{P}_\theta((X_1, ..., X_n) \in \mathcal{R})\} = \alpha\,.$$

On parle de **fonction de risque de 2nde espèce** définie sur Θ_1 pour

$$\theta \to \beta(\theta) = \mathbb{P}_\theta((X_1, ..., X_n) \notin \mathcal{R})\,.$$

On parle de **fonction puissance** définie sur Θ_1 pour

$$\theta \to 1 - \beta(\theta) = \mathbb{P}_\theta((X_1, ..., X_n) \in \mathcal{R})\,.$$

• **Test sur le coefficient de corrélation linéaire**:
Soit un vecteur gaussien (X, Y). On note $\rho = \mathrm{Cor}(X, Y) \in [-1, 1]$ le coefficient de corrélation linéaire entre X et Y. On souhaite tester au niveau α l'hypothèse nulle H_0: $\mathrm{Cor}(X, Y) = 0$ contre l'hypothèse alternative H_1: $\mathrm{Cor}(X, Y) \neq 0$. Soit $(X_1, Y_1), \ldots, (X_n, Y_n)$ un échantillon i.i.d. distribué comme le couple (X, Y). Soit l'estimateur suivant de $\rho = \mathrm{Cor}(X, Y)$ (dû à Pearson):

$$\widehat{\rho}_n = \frac{\sum_{i=1}^n \left(X_i - \overline{X}_n\right)\left(Y_i - \overline{Y}_n\right)}{\sqrt{\left(\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2\right)\left(\sum_{i=1}^n \left(Y_i - \overline{Y}_n\right)^2\right)}} \tag{1}$$

en notant $\overline{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ et $\overline{Y}_n = \dfrac{1}{n}\sum_{i=1}^{n} Y_i$. Soit la statistique de test:

$$T_n = \sqrt{n-2}\,\frac{\widehat{\rho}_n}{\sqrt{1-\widehat{\rho}_n^2}}.$$

Sous H_0, la statistique T_n suit la loi de Student à $(n-2)$ degrés de liberté $T(n-2)$. La région critique est:

$$\mathcal{R} = \{(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^{2n} : |t_n| > F^{-1}_{T(n-2)}(1-\alpha/2)\}$$

en notant $F^{-1}_{T(n-2)}(1-\alpha/2)$ le quantile d'ordre $(1-\alpha/2)$ de la loi $T(n-2)$.

La fonction `cor.test` du logiciel `R` implémente ce test.

• **Test d'association de Spearman**:
Soit un vecteur (X, Y) dont la fonction de répartition jointe est continue en ses deux arguments. On se demande s'il existe une forme d'association (corrélation non nécessairement de forme linéaire) entre X et Y. On souhaite tester au niveau α l'hypothèse nulle H_0: "il existe une association entre X et Y" contre l'hypothèse alternative H_1: "il n'existe pas d'association entre X et Y". Soit $(X_1, Y_1), \ldots, (X_n, Y_n)$ un échantillon i.i.d. distribué comme le couple (X, Y). Notons R_{X_i} le rang de X_i dans $(X_1, \ldots, X_n)$ et R_{Y_i} le rang de Y_i dans $(Y_1, \ldots, Y_n)$ Spearman propose de calculer un coefficient d'association à partir de la formule (1) de Pearson mais en remplaçant X_i par R_{X_i} et Y_i par R_{Y_i}. Remarquons que

$$\frac{1}{n}\sum_{i=1}^{n} R_{X_i} = \frac{n(n+1)}{2} = \frac{1}{n}\sum_{i=1}^{n} R_{Y_i}$$

. Cela donne

$$\widehat{r}_n = \frac{\sum_{i=1}^{n}\left(R_{X_i} - \frac{n(n+1)}{2}\right)\left(R_{Y_i} - \frac{n(n+1)}{2}\right)}{\sqrt{\left(\sum_{i=1}^{n}\left(R_{X_i} - \frac{n(n+1)}{2}\right)^2\right)\left(\sum_{i=1}^{n}\left(R_{Y_i} - \frac{n(n+1)}{2}\right)^2\right)}}.$$

Quelques calculs et le fait que

$$\frac{1}{n}\sum_{i=1}^{n} (R_{X_i})^2 = \frac{n(n+1)(2n+1)}{6} = \frac{1}{n}\sum_{i=1}^{n} (R_{Y_i})^2$$

amènent à l'écriture suivante, en notant $D_i = R_{X_i} - R_{Y_i}$,

$$\widehat{r}_n = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)}.$$

Sous H_0, la loi de T_n ne dépend que n. La région critique est:

$$\mathcal{R} = \{(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^{2n} : |\widehat{r}_n| > K_{\alpha,n}\}$$

en notant $K_{\alpha,n}$ la valeur telle que $\mathbb{P}_{H_0}(|\widehat{r}_n| > K_{\alpha,n}) = \alpha$ à déterminer numériquement ou à partir de la tabulation.

La fonction `cor.test` du logiciel `R` implémente ce test.
La fonction `spearman.test` du package `pspearman` du logiciel `R` implémente également ce test (en une version plus précise).

• **Test d'association de Kendall**:
Soit un vecteur (X, Y) dont la fonction de répartition jointe est continue en ses deux arguments. On se demande s'il existe une forme d'association (corrélation non nécessairement de forme linéaire) entre X et Y. On souhaite tester au niveau α l'hypothèse nulle H_0: "il existe une association entre X et Y" contre l'hypothèse alternative H_1: "il n'existe pas d'association entre X et Y". Soit $(X_1, Y_1), \ldots, (X_n, Y_n)$ un échantillon i.i.d. distribué comme le couple (X, Y). Notons

$$\text{sign} : x \to \begin{cases} 1 & \text{si } x > 0 \\ 0, & \text{si } x = 0 \\ 1, & \text{si } x < 0. \end{cases}$$

Kentall a proposé le coefficient d'association suivant:

$$\tau_n = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I(\text{sign}(X_i - X_j) = \text{sign}(Y_i - Y_j)) - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I(\text{sign}(X_i - X_j) = -\text{sign}(Y_i - Y_j))}{n(n-1)/2}.$$

Sous H_0, la loi de τ_n ne dépend que n. On peut montrer que, sous H_0, on a la convergence en loi suivante:

$$\frac{\tau_n}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1).$$

La région critique est:

$$\mathcal{R} = \{(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^{2n} : |\tau_n| > K_{\alpha,n}\}$$

en notant $K_{\alpha,n}$ la valeur telle que $\mathbb{P}_{H_0}(|\widehat{r}_n| > K_{\alpha,n}) = \alpha$ à déterminer numériquement ou à partir de la tabulation (ou de l'approximation normale).

La fonction `cor.test` du logiciel `R` implémente ce test.
La fonction `Kendall` du package `Kendall` du logiciel `R` implémente également ce test (en une version plus précise).

Exercice 1.

1. Illustrer de manière empirique à partir de données simulées le comportement de la taille du test.
2. Illustrer de manière empirique à partir de données simulées le comportement de la fonction puissance.

Vous veillerez à faire varier tour à tour:

- le risque de 1ère espèce α,
- la taille de l'échantillon n,
- la variabilité du phénomène,
- le cas échéant l'ampleur de l'écart à H_0.