Université de Strasbourg
M1 - Magistère
Statistique - études de cas

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

Sujet 12: Tests d'indépendance

• **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 au niveau α 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

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

• **Tests du chi-deux d'indépendance entre deux variables**:
Cas de deux lois discrètes:
Soit un couple $X = (Y, Z)$ de variables discrètes. On suppose que les variables Y et Z sont à valeurs respectivement dans $\mathcal{Y} = \{b_1, \ldots, b_J\}$ et $\mathcal{Z} = \{c_1, \ldots, c_L\}$. La loi de probabilité de X est donnée par $(p_{j,\ell})_{1 \le j \le J 1 \le \ell \le L}$ où $p_{j,\ell} = \mathbb{P}(Y = b_j, Z = c_\ell)$ pour $1 \le j \le J$ et $1 \le \ell \le L$ avec $\sum_{j=1}^{J} \sum_{\ell=1}^{L} p_{j,\ell} = 1$. Les lois marginales de Y et Z sont données respectivement par $(p_{j,.})_{1 \le j \le J}$ où $p_{j,.} = \mathbb{P}(Y = b_j)$ et $(p_{.,\ell})_{1 \le \ell \le L}$ où $p_{.,\ell} = \mathbb{P}(Z = c_\ell)$. On souhaite tester l'indépendance de

Y et Z au niveau α. Formulons alors l'hypothèse nulle en H_0: "Y et Z sont indépendantes" et l'hypothèse alternative H_1: "Y et Z ne sont pas indépendantes". L'indépendance de Y et Z est équivalente à l'égalité entre la loi jointe de (Y, Z) et le produit des lois marginales de Y et Z respectivement. On peut alors reformuler l'hypothèse nulle en H_0: $p_{j,\ell} = p_{j,.}p_{.,\ell}$ pour $j = 1, \ldots, J$ et $\ell = 1, \ldots, L$ et l'hypothèse alternative en H_1: $\exists(j_0, \ell_0)$ tel que $p_{j_0,\ell_0} \neq p_{j_0,.}p_{.,\ell_0}$. Soit un échantillon i.i.d. $(((Y_1, Z_1), \ldots, (Y_n, Z_n))$ distribué comme le couple $X = (Y, Z)$. Notons

$$N_{j,\ell} = \sum_{j=1}^{J} \sum_{\ell=1}^{L} I(Y_j = a_j, Z_\ell = c_\ell)$$

l'effectif observé dans la catégorie (j, ℓ),

$$N_{j,.} = \sum_{j=1}^{J} I(Y_j = a_j)$$

l'effectif cumulé observé dans la catégorie j et

$$N_{.,\ell} = \sum_{\ell=1}^{L} I(Z_\ell = c_\ell)$$

l'effectif cumulé observé dans la catégorie ℓ. L'EMV de $(p_{j,\ell})_{1\leq j\leq J, 1\leq \ell\leq L}$ dans le modèle multinomial sous-jacent pour $(N_{j,\ell})_{1\leq j\leq J, 1\leq \ell\leq L}$ est alors $(\widehat{p}_{j,\ell})_{1\leq j\leq J, 1\leq \ell\leq L}$ où $\widehat{p}_{j,\ell} = \dfrac{N_{j,\ell}}{n}$. L'EMV de $(p_{j,.})_{1\leq j\leq J}$ dans le modèle multinomial sous-jacent pour $(N_{j,.})_{1\leq j\leq J}$ est $(\widehat{p}_{j,.})_{1\leq j\leq J}$ où $\widehat{p}_{j,.} = \dfrac{N_{j,.}}{n}$. L'EMV de $(p_{.,\ell})_{1\leq \ell\leq L}$ dans le modèle multinomial sous-jacent pour $(N_{.,\ell})_{1\leq \ell\leq L}$ est $(\widehat{p}_{.,\ell})_{1\leq \ell\leq L}$ où $\widehat{p}_{.,\ell} = \dfrac{N_{.,\ell}}{n}$. Soit la statistique de test:

$$T_n = n \sum_{j=1}^{J} \sum_{\ell=1}^{L} \frac{(\widehat{p}_{j,\ell} - \widehat{p}_{j,.}\widehat{p}_{.,\ell})^2}{\widehat{p}_{j,.}\widehat{p}_{.,\ell}}\,.$$

Sous H_0, la statistique T_n converge en loi vers une variable de loi $\chi^2((J-1)(L-1))$. Sous H_1, la statistique T_n tend en probabilité vers ∞. La région critique associée au niveau asymptotique de test α est

$$\mathcal{R} = \{T_n > F^{-1}_{\chi^2((J-1)(L-1))}(1-\alpha)\}.$$

Cas d'une loi continue:
Ce qui précède s'applique si l'on discrétise le support de Y en une partition $\mathcal{Y} = \cup_{j=1}^{J} I_j$ et le support de Y en une partition $\mathcal{Z} = \cup_{\ell=1}^{L} I'_\ell$ de sorte que $p_{j,\ell} = \mathbb{P}(Y \in I_j, Z \in I'_\ell)$.

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

• **Tests de Hoeffding d'indépendance entre deux variables**:
Soit un couple $X = (Y, Z)$ de variables aléatoires à valeurs dans $\mathcal{X} = \mathcal{Y} \times \mathcal{Z}$ dont la fonction de répartition jointe $F_{Y,Z}(.,.)$ est continue en chacun de ses arguments. Notons F_Y la fonction de répartition marginale de Y et F_Z la fonction de répartition marginale de Z. On souhaite tester l'indépendance de Y et Z au niveau α. Formulons alors l'hypothèse nulle H_0: "Y et Z sont indépendantes" et l'hypothèse alternative H_1: "Y et Z ne sont pas indépendantes".

L'indépendance de Y et Z est équivalente à l'égalité entre la loi jointe de (Y, Z) et le produit des lois marginales de Y et Z respectivement. On peut alors reformuler l'hypothèse nulle en H_0: $F_{Y,Z} = F_Y F_Z$ et l'hypothèse alternative en H_1: $F_{Y,Z} \neq F_Y F_Z$. La statistique de test est une version empirique de la mesure d'écart à l'indépendance suivante:

$$\iint \left(F_{Y,Z}(y,z) - F_Y(y)F_Z(z)\right)^2 F_{Y,Z}(dy, dz).$$

Soit un échantillon i.i.d. $((Y_1, Z_1), \ldots, (Y_n, Z_n))$ avec $n \geq 5$, distribué comme le couple $X = (Y, Z)$. La statistique de test est:

$$T_n = \frac{A_n - 2(n-2)B_n + (n-2)(n-3)C_n}{n(n-1)(n-2)(n-3)(n-4)}$$

où, en notant R_{Y_i} le rang de Y_i dans $(Y_1, \ldots, Y_n)$ et R_{Z_i} le rang de Z_i dans $(Z_1, \ldots, Z_n)$,

$$A_n = \sum_{i=1}^{n} (R_{Y_i} - 1)(R_{Y_i} - 2)(R_{Z_i} - 1)(R_{Z_i} - 2)$$

$$B_n = \sum_{i=1}^{n} (R_{Y_i} - 2)(R_{Z_i} - 2)D_i$$

$$C_n = \sum_{i=1}^{n} D_i(D_i - 1)$$

$$D_i = \sum_{j=1}^{n} I(Y_i > Y_j)I(Z_i > Z_j), \quad i = 1, \ldots, n.$$

On peut montrer que la loi de T_n sous H_0 ne dépend que de n (elle ne dépend pas de $F_{Y,Z}$, ni de F_Y, ni de F_Z). Notons t_n la réalisation de T_n. A n fixé, la région critique exacte associée au niveau de test α est

$$\{(x_1, \ldots, x_n) \in \mathcal{X}^n : t_n > k_{\alpha,n}\}$$

où $k_{\alpha,n}$ est déterminé numériquement par ordinateur (ou à partir d'une tabulation).
La fonction `hoeffd` du package `Hmisc` du logiciel `R` implémente ce test.

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.