Université de Strasbourg
M1 - Magistère
Statistique - études de cas

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

Sujet 11: Test de comparaison de deux distributions (ou plus)

• **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 du χ^2 d'homogénéité entre J populations indépendantes**:
Cas d'une loi discrète:
Soit un espace fini $\mathcal{X} = \{a_1, \ldots, a_K\}$. Soient J populations indépendantes dont on extrait J échantillons indépendants $(X_1^{(j)}, ..., X_{n_j}^{(j)})_{j=1,...,J}$. Pour $j = 1, \ldots, J$, le $j$ème échantillon $(X_1^{(j)}, ..., X_{n_j}^{(j)})$ compte n_j variables indépendantes, toutes distribuées comme une variable $X^{(j)}$ à valeurs dans $\mathcal{X}$. Pour $j = 1, \ldots, J$, la loi de probabilité de $X^{(j)}$ est donnée par $(p_1^{(j)}, \ldots, p_K^{(j)})$ où $p_k^{(j)} = \mathbb{P}(X^{(j)} = a_k)$ pour $k = 1, \ldots, K$ avec $\sum_{k=1}^{K} p_k^{(j)} = 1$. Soit $(N_1^{(j)}, \ldots, N_K^{(j)}$ pour

$j \in \{1,\dots,J\}$ le vecteur des différents effectifs où $N_k^{(j)} = \sum_{i=1}^{n_j} I(X_i^{(j)} = a_k)$. Par définition, le vecteur $(N_1^{(j)},\dots,N_K^{(j)}$ suit une loi multinomiale de paramètres $(n_j, p_1^{(j)},\dots,p_K^{(j)})$ pour $j \in \{1,\dots,J\}$. L'EMV de $p_k^{(j)}$ dans ce modèle multinomial est $\widehat{p}_k^{(j)} = \dfrac{N_k^{(j)}}{n_j}$. On souhaite tester au niveau α si les lois de probabilité $(p_1^{(j)},\dots,p_K^{(j)})$ sont identiques pour $j = 1,\dots,J$. Formulons alors H_0: "les J lois de probabilité sont identiques" et H_1: "au moins l'une des J lois de probabilité diffèrent des autres". Notons $(p_1,\dots,p_K)$ la loi de probabilité commune sous H_0. Sous H_0, l'EMV de p_k est $\widehat{p}_k = \dfrac{\sum_{j=1}^J N_k^{(j)}}{\sum_{j=1}^J n_j}$. Soit la statistique de test (en notant $n = \sum_{j=1}^J n_j$):

$$T_n = \sum_{j=1}^{J}\sum_{k=1}^{K} \frac{(N_k^{(j)} - n_j\widehat{p}_k)^2}{n_j\widehat{p}_k}\,.$$

Sous H_0, la statistique T_n convergence en loi vers $\chi^2((J-1)(K-1))$ lorsque tous les n_j tendent vers ∞. Sous H_1, si tous les rapports n_j/n restent inférieurement bornés par une constante strictement positive (indépendante de n), alors T_n tend presque-sûrement vers ∞. Notons t_n la réalisation de T_n. La région critique associée au niveau asymptotique de test α est

$$\mathcal{R} = \{(x_1^{(j)},...,x_{n_j}^{(j)})_{j=1,...,J} : t_n > F^{-1}_{\chi^2((J-1)(K-1))}(1-\alpha)\}.$$

Cas d'une loi continue:
Ce qui précède peut s'appliquer si l'on discrétise le support des $X^{(j)}$ en une partition $\mathcal{X} = \cup_{k=1}^K I_k$ et de remplacer les a_k par les I_k de sorte que $p_k = \mathbb{P}(X^{(j)} \in I_k)$.

La fonction `prop.test` (package `stats` chargé par défaut lors du lancement du logiciel `R`) implémente ce test dans le cas d'une loi discrète à support fini.

• **Test de Kolmogorov-Smirnov à deux échantillons indépendants**:
Soit une variable X de loi F_X continue et soit une variable Y de loi F_Y également continue, indépendante de X. On veut comparer les deux distributions continues F_X et F_Y sur la base de deux échantillons indépendants. Formellement, on souhaite tester au niveau de risque de 1ère espèce α l'hypothèse H_0: $F_X = F_Y$ contre l'hypothèse H_1: $F_X \neq F_Y$.
Soit $(X_1,\dots,X_{n_X})$ un échantillon i.i.d. distribué comme X, de fonction de répartition F_X. Soit $(Y_1,\dots,Y_{n_Y})$ un échantillon i.i.d. distribué comme Y, de fonction de répartition F_Y, indépendant de $(X_1,\dots,X_{n_X})$. Soit $\widehat{F}_{X,n_X}$ la fonction de répartition empirique des X_i définie pour $x \in \mathbb{R}$ par

$$\widehat{F}_{X,n_X}(x) = \frac{1}{n_X}\sum_{i=1}^{n_X} I(X_i \leq x).$$

Soit $\widehat{F}_{Y,n_Y}$ la fonction de répartition empirique des Y_i définie pour $x \in \mathbb{R}$ par

$$\widehat{F}_{Y,n_Y}(x) = \frac{1}{n_Y}\sum_{i=1}^{n_Y} I(Y_i \leq x).$$

Soit la statistique de test:

$$D_{n_X,n_Y} = \sqrt{\frac{n_X n_Y}{n_X+n_Y}}\sup_{x\in\mathbb{R}} |\widehat{F}_{X,n_X}(x) - \widehat{F}_{Y,n_Y}(x)|\,.$$

Intuitivement, si H_0 est vraie, D_{n_X,n_Y} prend de petites valeurs et on peut montrer que sa loi ne dépend que de n_X et de n_Y. Notons d_{n_X,n_Y} la réalisation de D_{n_X,n_Y}.
La loi exacte de D_{n_X,n_Y} est tabulée pour les "petites" valeurs de n_X et n_Y. La région critique associée au niveau α est alors

$$\{(x_1,\ldots,x_{n_X},y_1,\ldots,y_{n_Y}) \in \mathbb{R}^{n_X+n_Y} : d_{n_X,n_Y} > k_{\alpha,n_X,n_Y}\}$$

où k_{α,n_X,n_Y} satisfait $\mathbb{P}_{F_X=F_Y}(D_{n_X,n_Y} > k_{\alpha,n_X,n_Y}) = \alpha$.
Smirnov (1939) a montré que, sous H_0, lorsque $n_X, n_Y \to \infty$, la statistique T_{n_X,n_Y} suit asymptotiquement la loi d'une variable Z de fonction de répartition donnée par $K(.)$ où

$$K(y) = \sum_{-\infty}^{\infty} (-1)^k \exp(-2k^2y^2) = 1 + 2\sum_{k=1}^{\infty} (-1)^k \exp(-2k^2y^2)\,.$$

On peut également obtenir la convergence presque-sûre de D_n vers ∞ sous H_1. La région critique associée au niveau asymptotique de test α est

$$\{(x_1,\ldots,x_{n_X},y_1,\ldots,y_{n_Y}) \in \mathbb{R}^{n_X+n_Y} : d_{n_X,n_Y} > k_{\alpha}\}$$

où k_α satisfait $K(k_\alpha) = 1-\alpha$.
La fonction `ks.test` (package `stats` chargé par défaut lors du lancement du logiciel `R`) implémente les versions exacte et asymptotique de ce test.

• **Test de Cramer-von-Mises à deux échantillons indépendants**:
Soit une variable X de loi F_X continue et soit une variable Y de loi F_Y également continue, indépendante de X. On veut comparer les deux distributions continues F_X et F_Y sur la base de deux échantillons indépendants. Formellement, on souhaite tester au niveau de risque de 1ère espèce α l'hypothèse H_0: $F_X = F_Y$ contre l'hypothèse H_1: $F_X \neq F_Y$.
Soit $(X_1,\ldots,X_{n_X})$ un échantillon i.i.d. distribué comme X, de fonction de répartition F_X. Soit $(Y_1,\ldots,Y_{n_Y})$ un échantillon i.i.d. distribué comme Y, de fonction de répartition F_Y, indépendant de $(X_1,\ldots,X_{n_X})$. Soit $\widehat{F}_{X,n_X}$ la fonction de répartition empirique des X_i définie pour $x \in \mathbb{R}$ par

$$\widehat{F}_{X,n_X}(x) = \frac{1}{n_X}\sum_{i=1}^{n_X} I(X_i \leq x).$$

Soit $\widehat{F}_{X,n_Y}$ la fonction de répartition empirique des Y_i définie pour $x \in \mathbb{R}$ par

$$\widehat{F}_{Y,n_Y}(x) = \frac{1}{n_Y}\sum_{i=1}^{n_Y} I(Y_i \leq x).$$

Introduisons également la fonction de répartition empirique de l'échantillon aggloméré:

$$\widehat{F}_{n_X+n_Y}(x) = \frac{1}{n_X+n_Y}\left(\sum_{i=1}^{n_X} I(X_i \leq x) + \sum_{i=1}^{n_Y} I(Y_i \leq x)\right).$$

Soit la statistique de test:

$$D_{n_X,n_Y} = \frac{n_X n_Y}{n_X+n_Y}\int \left(\widehat{F}_{X,n_X}(x) - \widehat{F}_{Y,n_Y}(x)\right)^2 d\widehat{F}_{n_X+n_Y}(x).$$

Intuitivement, si H_0 est vraie, D_{n_X,n_Y} prend de petites valeurs et on peut montrer que sa loi ne dépend que de n_X et de n_Y.

Notons d_{n_X,n_Y} la réalisation de D_{n_X,n_Y}. La région critique associée au niveau α est alors

$$\{(x_1,\ldots,x_{n_X},y_1,\ldots,y_{n_Y}) \in \mathbb{R}^{n_X+n_Y} : d_{n_X,n_Y} > k_{\alpha,n_X,n_Y}\}$$

où k_{α,n_X,n_Y} satisfait $\mathbb{P}_{F_X=F_Y}(D_{n_X,n_Y} > k_{\alpha,n_X,n_Y}) = \alpha$.
La fonction `cvmts.test` du package `CvM2SL2Test` 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.
3. Tests paramétriques: Illustrer de manière empirique à partir de données simulées la robustesse ou au contraire la sensibilité du test aux hypothèses sous-jacentes
4. Tests non-paramétriques: faire varier la loi servant à générer les données ainsi que la valeur de son/ses paramètre(s)

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.