Université de Strasbourg
M1 - Magistère
Statistique - études de cas

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

Sujet 6: Tests d'adéquation à une loi donnée (avec paramètres également donnés)

• **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'adéquation à une loi donnée**:
Cas d'une loi discrète:
Soit une variable aléatoire discrète X à valeurs dans un espace fini $\mathcal{X} = \{a_1, \ldots, a_K\}$. La loi de probabilité de X est donnée par $(p_1, \ldots, p_K)$ où $p_k = \mathbb{P}(X = a_k) \in]0, 1[$ pour $k = 1, \ldots, K$ avec $\displaystyle\sum_{k=1}^{K} p_k = 1$. Soit $(p_{1,0}, \ldots, p_{K,0}) \in]0, 1[^K$ avec $\displaystyle\sum_{k=1}^{K} p_{k,0} = 1$ une loi de probabilité discrète fixée. On souhaite tester l'hypothèse nulle H_0: $(p_1, \ldots, p_K) = (p_{1,0}, \ldots, p_{K,0})$ contre l'hypothèse aletrnative H_1: $(p_1, \ldots, p_K) \neq (p_{1,0}, \ldots, p_{K,0})$ au niveau α.

Soit un échantillon i.i.d. $(X_1, ..., X_n)$ distribué comme X. Notons $N_k = \sum_{i=1}^{n} I(X_i = a_k)$ l'effectif observé dans la catégorie k. Par définition, le vecteur $(N_1, \ldots, N_K)$ suit la loi multinomiale de paramètre $(n, p_1, \ldots, p_K)$ de loi donnée par

$$\mathbb{P}(N_1 = n_1, \ldots, N_K = n_K) = \frac{n!}{n_1! \ldots n_K!} p_1^{n_1} \ldots p_K^{n_K}\,.$$

L'EMV de $(p_1, \ldots, p_K)$ dans le modèle multinomial sous-jacent pour $(N_1, ..., N_K)$ est alors $(\widehat{p}_1, \ldots, \widehat{p}_K)$ où $\widehat{p}_k = \frac{N_k}{n}$. Soit la statistique de test:

$$T_n = n \sum_{k=1}^{K} \frac{(\widehat{p}_k - p_{k,0})^2}{p_{k,0}} = \sum_{k=1}^{K} \frac{(N_k - np_{k,0})^2}{np_{k,0}}$$

qui mesure l'écart aléatoire entre les effectifs observés et les effectifs espérés sous H_0. Sous H_0, la statistique T_n converge en loi vers une variable de loi $\chi^2(K-1)$. En pratique toutefois, il est recommandé de n'utiliser cette approximation en loi que si n est suffisamment grand pour que $n \min(p_{1,0}, ..., p_{K,0}) \geq 5$. Sous H_1, la statistique 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

$$\mathcal{R} = \{(x_1, \ldots, x_n) \in \{a_1, \ldots, a_K\}^n : t_n > F^{-1}_{\chi^2(K-1)}(1-\alpha)\}.$$

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

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

• **Test de Kolmogorov-Smirnov de conformité à une loi continue**:
Soit une variable X de fonction de répartition F_X de support $\mathcal{X}$. On souhaite tester l'ajustement à une distribution continue fournie par l'utilisateur et entièrement spécifiée, à savoir, ses éventuels paramètres sont également fournis par l'utilisateur. Formellement, on souhaite tester H_0: $F = F_0$ contre H_1: $F \neq F_0$ au niveau α, avec F_0 continue.
Soit $(X_1, \ldots, X_n)$ un échantillon i.i.d. de fonction de répartition F_X. Soit $\widehat{F}_n$ la fonction de répartition empirique des X_i définie pour $x \in \mathbb{R}$ par $\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$. Soit la statistique de test:

$$D_n = \sqrt{n} \sup_{x \in \mathbb{R}} |\widehat{F}_n(x) - F_0(x)|\,.$$

Notons d_n la réalisation de D_n.
La loi de D_n sous H_0 ne dépend que de n (elle ne dépend pas de F_0), ce qui permet d'obtenir une version exacte du test. A n fixé, la région critique exacte associée au niveau de test α est

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

où $k_{\alpha,n}$ est déterminé numériquement par ordinateur (ou à partir d'une tabulation).
Kolmogorov (1933) a montré que $D_n \xrightarrow{\mathcal{D}} Z$ où Z est une variable aléatoire dont la fonction de répartition notée K est donnée par

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

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

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

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 de conformité**:
Soit une variable X de fonction de répartition F_X de support $\mathcal{X}$. On souhaite tester l'ajustement à une distribution continue fournie par l'utilisateur et entièrement spécifiée (à savoir, ses éventuels paramètres sont également fournis par l'utilisateur). Formellement, on souhaite tester H_0: $F = F_0$ contre H_1: $F \neq F_0$ au niveau α, avec F_0 continue.
Soit $(X_1, \ldots, X_n)$ un échantillon i.i.d. de fonction de répartition F_X. Soit $\widehat{F}_n$ la fonction de répartition empirique des X_i définie pour $x \in \mathbb{R}$ par $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$. Soit la statistique de test:

$$D_n = n \int_{-\infty}^{\infty} \left(\widehat{F}_n(x) - F_0(x)\right)^2 dF_0(x).$$

La loi de D_n sous H_0 ne dépend que de n (elle ne dépend pas de F_0). Notons d_n la réalisation de D_n. A n fixé, la région critique associée au niveau de test α est

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

où $k_{\alpha,n}$ est approximé numériquement par ordinateur (ou à partir d'une tabulation).
La fonction `cvm.test` du package `goftest` 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.

Dans le cas où le test de conformité du χ^2 est utilisé dans le cas d'une variable continue, quelle est la sensibilité au choix du découpage en classes?