Université de Strasbourg
M1 - Magistère
Statistique - études de cas

Ségolen Geffray
geffray@math.unistra.fr
Année 2020/2021

TP1: tests d'adéquation et d'indépendance

• Test du χ^2 d'adéquation à une loi donnée (avec paramètre(s) également donné(s)): 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 $\sum_{k=1}^{K} p_k = 1$. Soit $(p_{1,0}, \ldots, p_{K,0}) \in]0,1[^K$ avec $\sum_{k=1}^{K} p_{k,0} = 1$ 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 alternative 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 $N_k = \sum_{i=1}^{n} I(X_i = a_k)$. Soit la statistique de test:

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

Sous H_0, la statistique T_n converge en loi vers une variable de loi $\chi^2(K-1)$ lorsque $n \to \infty$. 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 ∞ lorsque $n \to \infty$. Notons t_n la réalisation de T_n. La région de rejet associée au niveau asymptotique de risque de 1ère espèce α est

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

La p-valeur du test est $\mathbb{P}(\chi^2(K-1) > t_n)$ (avec un léger abus de notation).
La fonction `chisq.test` du logiciel `R` implémente ce test.

• Test de Kolmogorov-Smirnov d'adéquation 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

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

où $k_{\alpha,n}$ est déterminé numériquement par ordinateur (ou à partir d'une tabulation).
Kolmogorov (1933) a montré que $D_n \overset{\mathcal{D}}{\to} Z$ 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 de rejet associée au niveau asymptotique de test α est

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

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 d'adéquation**:
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 $(X_1, \dots, X_n)$ 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 $\widehat{F}_n(x) = \dfrac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$. Soit la statistique de test D_n qui est une mesure de l'écart entre une fonction de répartition théorique et une fonction de répartition empirique:

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

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 de rejet associée au niveau de test α est

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

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.

• **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 $(p_{j,\ell})_{1\leq j\leq J 1\leq \ell\leq L}$ où $p_{j,\ell} = \mathbb{P}(Y = b_j, Z = c_\ell)$ pour $1 \leq j \leq J$ et $1 \leq \ell \leq L$ avec $\sum_{j=1}^{J}\sum_{\ell=1}^{L} p_{j,\ell} = 1$. Les lois marginales de Y et Z sont données respectivement par $(p_{j,.})_{1\leq j\leq J}$ où $p_{j,.} = \mathbb{P}(Y = b_j)$ et $(p_{.,\ell})_{1\leq \ell\leq L}$ 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

$$N_{j,\ell} = \sum_{i=1}^{n} I(Y_i = a_j, Z_i = c_\ell)$$

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

$$N_{j,.} = \sum_{i=1}^{n} I(Y_i = a_j)$$

l'effectif cumulé observé dans la catégorie j et

$$N_{.,\ell} = \sum_{i=1}^{n} I(Z_i = c_\ell)$$

l'effectif cumulé observé dans la catégorie ℓ. 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 $\chi^2((J-1)(L-1))$ lorsque $n \to \infty$. En pratique toutefois, il est recommandé de n'utiliser l'approximation en loi que si n est suffisamment grand pour que $n \min_{1\leq j\leq J, 1\leq \ell\leq L}\{p_{j,.}p_{.,\ell}\} \geq 5$ ou 10 selon les auteurs. Sous H_1, la statistique T_n tend en probabilité vers ∞ lorsque $n \to \infty$. La région critique associée au niveau asymptotique de test α est

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

La p-valeur du test est $\mathbb{P}(\chi^2((J-1)(L-1)) > t_n)$ (avec un léger abus de notation).

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

• Illustration numérique:

Elaborer un plan de simulations explicitant le nombre d'échantillons générés noté M, la taille requise pour l'échantillon notée n et les différentes lois utilisées pour simuler les échantillons. Le travail à effectuer peut se mettre sous la forme de l'algorithme suivant:

• fixer m_0,
• choisir une loi de probabilité aux paramètres près, ce qui induit un choix de $\mathbb{E}[X]$,
• pour $n = 20, 50, 100, 200, 500$,

- pour $m = 1, ..., M$,
  - simuler un échantillon i.i.d. $\left(X_1^{\{m\}}, ..., X_n^{\{m\}}\right)$ selon la loi choisie,
  - déterminer si le test accepte ou rejette l'hypothèse nulle sur la base de l'échantillon simulé,
- déterminer la proportion empirique d'erreurs de $1^{\text{ère}}$ espèce ou la puissance empirique, selon le cas considéré.