Université de Strasbourg
M1 - Magistère
Statistique - études de cas

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

Sujet 7: Tests d'adéquation à une famille de lois (ou test d'adéquation à une loi avec paramètres estimés par maximum de vraisemblance)

• **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 avec paramètres estimés par maximum de vraisemblance**:

Cas d'une loi discrète (sans regrouper les valeurs possibles en classes):
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

$\sum_{k=1}^{K} p_k = 1$. On souhaite tester au niveau de risque de 1ère espèce α si X est issue d'une famille discrète paramétrée par $\lambda \in \Lambda$ où Λ est un ouvert non vide de $\mathbb{R}^d$ avec $d < K-1$. Notons p_λ la loi de probabilité de X lorsque $\lambda \in \Lambda$ est la valeur du paramètre ie $p_\lambda = (p_{1,\lambda}, \ldots, p_{K,\lambda})$ où $p_{k,\lambda} = \mathbb{P}_\lambda(X = a_k) \in]0,1[$ pour $k = 1, \ldots, K$ avec $\sum_{k=1}^{K} p_{k,\lambda} = 1$. Formulons alors l'hypothèse nulle H_0: "adéquation à la famille de lois donnée" (ie pour tout $k = 1, \ldots, K$, $p_k = p_{k,\lambda}$ pour un $\lambda \in \Lambda$) et l'hypothèse alternative H_1: "non-adéquation à la famille de lois donnée".
Soit un échantillon i.i.d. $(X_1, ..., X_n)$ distribué comme X. Notons $\widehat{p}_k = \dfrac{N_k}{n}$ avec

$$N_k = \sum_{i=1}^{n} I(X_i = a_k).$$

Soit $\widehat{\lambda}_n$ l'EMV de λ dans le modèle postulé sous H_0. Soit la statistique de test:

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

Sous H_0, la statistique T_n converge en loi vers la loi $\chi^2(K-d-1)$ pourvu que le modèle postulé soit régulier. Sous H_1, la statistique T_n tend en probabilité vers ∞. Notons t_n la réalisation de T_n. La région critique associée au niveau asymptotique de test α est

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

Cas d'une loi continue ou de données groupées en classes:
Ce qui précède peut s'appliquer dans les deux cas suivants:

1. Cas d'une variable X continue: on discrétise le support de la variable continue X en une partition finie $\mathcal{X} = \cup_{k=1}^{K} I_k$. Souvent, on choisit les intervalles I_k de sorte qu'ils sont équiprobables sous H_0. On remplace alors les a_k par les intervalles I_k de sorte que $p_k = \mathbb{P}(X \in I_k)$. On discrétise également la loi continue $(P_\lambda)_{\lambda\in\Lambda}$ postulée. On définit pour cela les $p_{\lambda,k} = \mathbb{P}_\lambda(X \in I_k) = \displaystyle\int_{I_k} dP_\lambda$ pour $k = 1, \ldots, K$. Les hypothèses du problème de test deviennent H_0: $p_k = p_{k,\lambda}$, pour tout $k = 1, \ldots, K$, pour un $\lambda \in \Lambda$ (ce qui n'équivaut pas à l'adéquation de la loi de X au modèle $(P_\lambda)_{\lambda\in\Lambda}$ contre H_1: $(p_1, \ldots, p_K) \neq (p_{1,\lambda}, \ldots, p_{K,\lambda})$ pour tout $\lambda \in \Lambda$. Soit $\widehat{\lambda}_n$ l'EMV de λ dans le modèle postulé sous H_0. Attention, cet EMV ne coïncide pas forcément avec l'EMV dans le modèle $(P_\lambda)_{\lambda\in\Lambda}$! On forme alors la statistique de test et on conclut comme précédemment, pourvu que la fonction $\lambda \to p_{k,\lambda}$ soit c^2 sur Λ pour $k = 1, \ldots, K$.
   Par exemple, si l'on souhaite tester l'adéquation à la loi exponentielle (avec paramètre λ estimé à partir des données et non arbitrairement fixé), on définit une partition finie de $\mathbb{R}^+ = \cup_{k=1}^{K} I_k$ et on pose

$$p_{k,\lambda} = \int_{I_k} \lambda \exp(-\lambda x) dx.$$

   Comme la loi sous H_0 du vecteur $(N_1, \ldots, N_K)$ en notant $N_k = \displaystyle\sum_{i=1}^{n} I(X_i \in I_k)$ est une loi multinomiale de paramètres $(n, p_{1,\lambda}, \ldots, p_{K,\lambda})$, l'EMV de λ sous H_0 s'obtient en

maximisant

$$\sum_{k=1}^{K} N_k p_{k,\lambda}.$$

2. Cas d'une variable X discrète dont on groupe certaines modalités en classes. Généralement, on a recours à cela pour qu'aucun des effectifs N_k, pour $k = 1, \ldots, K$, ne prenne une valeur trop petite et lorsque le support de la loi discrète est infini. On regroupe certaines valeurs possibles de X en une (nouvelle) partition finie du support de X, disons $\mathcal{X} = \cup_{k=1}^{K} I_k$. On remplace alors les a_k par les classes I_k de sorte que $p_k = \mathbb{P}(X \in I_k)$. On modifie également la loi discrète de départ $(P_\lambda)_{\lambda\in\Lambda}$ postulée. On définit pour cela les $p_{\lambda,k} = \mathbb{P}_\lambda(X \in I_k) = \displaystyle\int_{I_k} dP_\lambda$ pour $k = 1, \ldots, K$. Les hypothèses du problème de test deviennent H_0: $p_k = p_{k,\lambda}$, pour tout $k = 1, \ldots, K$, pour un $\lambda \in \Lambda$ (ce qui n'équivaut pas à l'adéquation de la loi de X au modèle $(P_\lambda)_{\lambda\in\Lambda}$ contre H_1: $(p_1, \ldots, p_K) \neq (p_{1,\lambda}, \ldots, p_{K,\lambda})$ pour tout $\lambda \in \Lambda$. Soit $\widehat{\lambda}_n$ l'EMV de λ dans le modèle postulé sous H_0. Attention, cet EMV ne coïncide pas forcément avec l'EMV dans le modèle $(P_\lambda)_{\lambda\in\Lambda}$! On forme alors la statistique de test et on conclut comme précédemment, pourvu que la fonction $\lambda \to p_{k,\lambda}$ soit c^2 pour $k = 1, \ldots, K$.

La fonction `goodfit` du package `vcd` du logiciel `R` implémente ce test pour les lois binomiale, de Poisson et binômiale négative, sans pratiquer de regroupement des données... autre que le regroupement des valeurs possibles au-delà de la plus grande observation avec cette dernière. Par exemple, si les observations sont

valeur	0	1	2	3	4	5	6	7
effectif	29	59	48	39	14	7	3	1

alors, la fonction `goodfit` calcule la statistique de test avec les classes suivantes

valeur	0	1	2	3	4	5	6	7 ou plus
effectif	29	59	48	39	14	7	3	1

De plus, l'attention de l'utilisateur est (modérément) attirée sur le caractère ad-hoc de ce test. La fonction `pearson.test` du package `nortest` implémente ce test pour la loi normale... en mettant en garde sur les limites à l'utilisation de ce test!

• **Test d'adéquation de Shapiro-Wilk à la loi gaussienne**:
Soit $(X_1, \ldots, X_n)$ un échantillon i.i.d. distribué comme une variable X. On souhaite tester au niveau α si la loi de X est la loi gaussienne notée $\mathcal{N}(\mathbb{E}[X], \mathrm{Var}(X))$. On formule alors l'hypothèse nulle H_0: la loi de X est gaussienne $\mathcal{N}(\mathbb{E}[X], \mathrm{Var}(X))$ et l'hypothèse alternative H_1: la loi de X n'est pas gaussienne. L'idée consiste à comparer deux statistiques $T_{n,1}$ et $T_{n,2}$ qui, sous l'hypothèse de normalité, estiment toutes deux $\mathrm{Var}(X)$ et, sous l'alternative, estiment des quantités différentes. Ces deux statistiques sont

$$T_1 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\overline{X}_n - X_i\right)^2$$

avec $\overline{X}_n = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} X_i$ et

$$T_2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} a_{i,n} X_{i,n} \right)^2$$

en notant $X_{1,n} < \ldots < X_{n,n}$ l'échantillon ordonné dans le sens croissant. Pour $i = 1, \ldots, n$, la constante $a_{i,n}$ est donnée par

$$(a_{1,n}, \ldots, a_{n,n}) = \frac{m^\top V^{-1}}{(m^\top V^{-1} V^{-1} m)^{1/2}}$$

où $m = (m_1, \ldots, m_n)^\top$ avec $m_i = \mathbb{E}[Z_{i,n}]$ où $Z_{1,n} < \ldots < Z_{n,n}$ est la version presque-sûrement ordonnée dans le sens croissant (dite statistique d'ordre) d'un échantillon i.i.d. $(Z_1, \ldots, Z_n)$ issu de la loi $\mathcal{N}(0,1)$, et V est la matrice de variance-covariance de ces $Z_{1,n} < \ldots < Z_{n,n}$. Soit la statistique de test:

$$T_n = \frac{T_{2,n}}{T_{1,n}}$$

dont la loi sous H_0 ne dépend que de n et de α. Sous H_0, la statistique T_n peut être interprétée comme le carré du coefficient de corrélation entre la série des quantiles générés à partir de la loi normale et les quantiles empiriques obtenus à partir des données. L'hypothèse nulle est donc peu vraisemblable lorsque T_n est trop petit. La région critique est alors

$$\mathcal{R} = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : t_n < k_{\alpha,n}\}$$

avec $\mathbb{P}_{\mathcal{N}(\mathbb{E}[X], \mathrm{Var}(X))}(T_n < k_{\alpha,n}) = \alpha$. Le seuil $k_{\alpha,n}$ est lu dans la table fournie par Shapiro et Wilk en fonction de n et α ou bien est déterminé par ordinateur.
La fonction `shapiro.test` (package `stats` chargé par défaut) du logiciel `R` implémente ce test. Le calcul est exact pour $n = 3$ et utilise des approximations sinon, différentes selon que $4 \leq n \leq 11$ ou $12 \leq n \leq 5000$.

• **Test d'adéquation de Lilliefors à la loi gaussienne**:
Soit $(X_1, \ldots, X_n)$ un échantillon i.i.d. distribué comme une variable X. On souhaite tester au niveau α si la loi de X est gaussienne $\mathcal{N}(\mathbb{E}[X], \mathrm{Var}(X))$. On formule alors l'hypothèse nulle H_0: la loi de X est gaussienne $\mathcal{N}(\mathbb{E}[X], \mathrm{Var}(X))$ et l'hypothèse alternative H_1: la loi de X n'est pas gaussienne. Notons $\overline{X}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$ un estimateur de $\mathbb{E}[X]$ et $S_n'^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n} \left(\overline{X}_n - X_i\right)^2$ un estimateur de $\mathrm{Var}(X)$. Définissons, pour $i = 1, ..., n$, les variables $Z_i = (X_i - \overline{X}_n)/\sqrt{S_n'^2}$. Soit $\widehat{F}_{Z,n}$ la fonction de répartition empirique des Z_i définie pour $x \in \mathbb{R}$ par

$$\widehat{F}_{Z,n}(x) = \frac{1}{n}\sum_{i=1}^{n} I(Z_i \leq x).$$

Notons ϕ la fonction de répartition de la loi $\mathcal{N}(0,1)$. Soit la statistique de test suivante qui mesure l'écart aléatoire entre la fonction de répartition empirique calculée à partir des données centrées réduites et la fonction de répartition théorique de la gaussienne centrée réduite:

$$T_n = \sup_{x \in \mathbb{R}} \left| \widehat{F}_{Z,n}(x) - \phi(x) \right|$$

dont la loi sous H_0 ne dépend que de n et de α. La région critique est alors

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

avec $\mathbb{P}_{\mathcal{N}(\mathbb{E}[X], \mathrm{Var}(X))}(T_n > k_{\alpha,n}) = \alpha$. Le seuil $k_{\alpha,n}$ est lu dans la table fournie par Shapiro et Wilk en fonction de n et α ou bien est déterminé par ordinateur.

La fonction `lillie.test` du package `nortest` du logiciel `R` implémente ce test.
NB: Le test de Lilliefors est une variante du test de Kolmogorov-Smirnov où les paramètres de la loi gaussienne sont estimés à partir des données.

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.

Dans le cas où le test du χ^2 d'adéquation avec paramètres estimés est utilisé dans le cas d'une variable continue, quelle est la sensibilité au choix du découpage en classes?
Que pensez-vous de la fiabilité du test du χ^2 avec paramètres estimés?