Université de Strasbourg
M1 - Magistère
Statistique - études de cas

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

Sujet 2: Tests de conformité sur le paramètre d'une loi de Bernoulli

---

• **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

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

• On souhaite ici comparer le paramètre d'une loi de Bernoulli à une norme fixée à partir d'un échantillon i.i.d. Soit $(X_1, \ldots, X_n)$ un échantillon i.i.d. de loi $\mathcal{B}(p)$ où $p \in]0, 1[$. Formellement, on souhaite tester l'hypothèse nulle H_0: $p = p_0$ contre l'hypothèse alternative H_1: $p \neq p_0$ au niveau α.

• **Test asymptotique n°1**:
Soit $\widehat{p}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$ l'EMV de p. Soit la statistique de test:

$$T_n = \sqrt{n}\frac{\widehat{p}_n - p_0}{\sqrt{p_0(1-p_0)}}\,.$$

Notons t_n sa réalisation. La région critique est

$$\mathcal{R} = \{(x_1, \ldots, x_n) \in \{0,1\}^n : |t_n| > F^{-1}_{\mathcal{N}(0,1)}(1-\alpha/2)\}$$

en notant $F^{-1}_{\mathcal{N}(0,1)}(1-\alpha/2)$ le fractile d'ordre $(1-\alpha/2)$ de la loi gaussienne standard $\mathcal{N}(0,1)$.
En pratique, il est recommandé que n soit assez grand pour que $np_0 > 5$ et $n(1-p_0) > 5$.
La fonction `prop.test` du logiciel `R` implémente ce test.

• **Test asymptotique n°2**:
On peut montrer que, sous H_0: $p = p_0$, on a

$$2\arcsin(\sqrt{\widehat{p}_n}) \sim \mathcal{N}\left(2\arcsin(\sqrt{p_0}), \frac{1}{\sqrt{n}}\right) .$$

Notons $\psi : x \to 2\arcsin(\sqrt{x})$. Cette fonction est strictement croissante d'inverse égale à $\psi^{-1} : y \to \sin^2(y/2)$. On recherche une région critique de la forme

$$\mathcal{R} = \left\{(x_1, \ldots, x_n) \in \mathbb{R}^n : \widehat{p}_n < k_\alpha \text{ ou } \widehat{p}_n > K_\alpha\right\}$$

pour $0 \leq k_\alpha < K_\alpha \leq 1$. En répartissant le risque de 1ère espèce de façon symétrique, il vient:

$$k_\alpha = \psi^{-1}\left(\psi(p_0) + \frac{F^{-1}_{\mathcal{N}(0,1)}(\alpha/2)}{\sqrt{n}}\right) \text{ et } K_\alpha = \psi^{-1}\left(\psi(p_0) + \frac{F^{-1}_{\mathcal{N}(0,1)}(1-\alpha/2)}{\sqrt{n}}\right) .$$

• **Test exact**:
On sait que, sous H_0: $p = p_0$, on a $\sum_{i=1}^n X_i \sim \mathcal{B}(n, p_0)$. On cherche le plus grand $k_{\alpha/2} \in \mathbb{N}$ tel que

$$\mathbb{P}_{p_0}\left(\sum_{i=1}^n X_i < k_{\alpha/2}\right) \leq \alpha/2$$

et le plus petit $K_{1-\alpha/2} \in \mathbb{N}$ tel que

$$\mathbb{P}_{p_0}\left(\sum_{i=1}^n X_i > K_{1-\alpha/2}\right) \leq \alpha/2.$$

La région critique est alors

$$\mathcal{R} = \left\{(x_1, \ldots, x_n) \in \mathbb{N}^n : \sum_{i=1}^n x_i \in \left(\cup_{k=0}^{k_{\alpha/2}-1}\{k\}\right) \cup \left(\cup_{k=K_{1-\alpha/2}+1}^{n}\{k\}\right)\right\}$$

avec la convention $\cup_{k=0}^{k_{\alpha/2}-1}\{k\} = \emptyset$ si $k_{\alpha/2} = 0$ et $\cup_{k=K_{1-\alpha/2}}^{n}\{k\} = \emptyset$ si $K_{1-\alpha/2} = n$.
La fonction `binom.test` du logiciel `R` implémente ce test.

• **Test du chi-deux de conformité pour variables discrètes**:
D'une manière générale, 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 $N_k = \sum_{i=1}^{n} I(X_i = a_k)$ l'effectif observé dans la catégorie k. Par définition, le vecteur $(N_1, \ldots, N_K)$ suit la loi multinomiale de paramètres $(n, p_1, \ldots, p_K)$ 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 $\widehat{p}_k = \dfrac{N_k}{n}$. Soit la statistique de test

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

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

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

La fonction `chisq.test` du logiciel `R` implémente ce test.
NB: dans le cadre d'un test d'adéquation à la loi de Bernoulli, notons que $K = 2$.

• **Test fondé sur une borne exponentielle**:
Le théorème des grandes déviations permet d'obtenir les deux inégalités suivantes:

$$\forall 1 > \delta > p_0, \quad \mathbb{P}_{p_0} \left(\frac{1}{n} \sum_{i=1}^{n} X_i \geq \delta \right) \leq \exp \left(-n \left(\delta \log \left(\frac{\delta}{p_0} \right) + (1-\delta) \log \left(\frac{1-\delta}{1-p_0} \right) \right) \right)$$

et

$$\forall 0 < \delta < p_0, \quad \mathbb{P}_{p_0} \left(\frac{1}{n} \sum_{i=1}^{n} X_i \leq \delta \right) \leq \exp \left(-n \left(\delta \log \left(\frac{\delta}{p_0} \right) + (1-\delta) \log \left(\frac{1-\delta}{1-p_0} \right) \right) \right) .$$

Notons $k_{\alpha/2,n}$ et $K_{1-\alpha/2,n}$ les deux solutions, respectivement inférieure et supérieure à p_0, de l'équation en δ:

$$\delta \log \left(\frac{\delta}{p_0} \right) + (1-\delta) \log \left(\frac{1-\delta}{1-p_0} \right) = \frac{1}{n} \log \left(\frac{2}{\alpha} \right) .$$

La région critique est alors

$$\mathcal{R} = \left\{ (x_1, \ldots, x_n) \in \mathbb{N}^n : \frac{1}{n} \sum_{i=1}^{n} x_i < k_{\alpha/2,n} \text{ ou } \frac{1}{n} \sum_{i=1}^{n} x_i > K_{1-\alpha/2,n} \right\} .$$

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. Illustrer de manière empirique à partir de données simulées la robustesse ou au contraire la sensibilité du test aux hypothèses sous-jacentes.

Vous veillerez à faire varier tour à tour:

- le risque de 1ère espèce α,
- la taille de l'échantillon n,
- la variabilité du phénomène.