Université de Strasbourg
M1 - Magistère
Statistique - études de cas

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

Sujet 8: Tests de comparaison de paramètres d'une loi de Bernoulli à partir de deux échantillons indépendants (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

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

• **Tests de comparaison de paramètres d'une loi de Bernoulli à deux échantillons indépendants**:
Soit $(X_1, \ldots, X_{n_X})$ un échantillon i.i.d. de loi $\mathcal{B}(p_X)$ où $p_X \in]0, 1[$. Soit $(Y_1, \ldots, Y_{n_Y})$ un échantillon i.i.d. de loi $\mathcal{B}(p_Y)$ où $p_Y \in]0, 1[$. On souhaite tester au niveau α l'hypothèse nulle H_0: $p_X = p_Y$ contre l'hypothèse alternative H_1: $p_X \neq p_Y$. Soit

$$\widehat{p}_{X,n_X} = \frac{1}{n_X} \sum_{i=1}^{n_X} X_i$$

un estimateur de p_X et soit

$$\widehat{p}_{Y,n_Y} = \frac{1}{n_Y}\sum_{i=1}^{n_Y} Y_i$$

un estimateur de p_Y. Notons également

$$\widehat{p}_{n_X+n_Y} = \frac{1}{n_X+n_Y}\left(\sum_{i=1}^{n_X} X_i + \sum_{i=1}^{n_Y} Y_i\right) = \frac{n_X\widehat{p}_{X,n_X} + n_Y\widehat{p}_{Y,n_Y}}{n_X+n_Y}.$$

Test asymptotique n°1:
Soit la statistique de test:

$$T_{n_X,n_Y} = \sqrt{\frac{n_X n_Y}{n_X+n_Y}}\frac{\widehat{p}_{X,n_X} - \widehat{p}_{Y,n_Y}}{\sqrt{\widehat{p}_{n_X+n_Y}(1-\widehat{p}_{n_X+n_Y})}}.$$

Sous H_0, la statistique T_n converge en loi vers une variable de loi $\mathcal{N}(0,1)$. En pratique toutefois, il est recommandé de n'utiliser cette approximation en loi que si n est suffisamment grand pour que $n_X p_X > 5$, $n_X(1-p_X) > 5$, $n_Y p_Y > 5$ et $n_Y(1-p_Y) > 5$. Sous H_1, la statistique T_n tend presque-sûrement vers ∞. La région critique est:

$$\mathcal{R} = \{(x_1,\ldots,x_{n_X}) \in \{0,1\}^{n_X}, (y_1,\ldots,y_{n_Y}) \in \{0,1\}^{n_Y} : |t_{n_X,n_Y}| > F^{-1}_{\mathcal{N}(0,1)}(1-\alpha/2)\}.$$

Test asymptotique n°2:
On peut montrer que

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

Notons $\psi : x \to 2\arcsin(\sqrt{x})$. Soit la statistique de test:

$$T_{n_X,n_Y} = \sqrt{\frac{n_X n_Y}{n_X+n_Y}}|\psi(\widehat{p}_{X,n_X}) - \psi(\widehat{p}_{Y,n_Y})|.$$

Sous H_0, la statistique T_n converge en loi vers une variable de loi $\mathcal{N}(0,1)$. Sous H_1, la statistique T_n tend presque-sûrement vers ∞. La région critique est:

$$\mathcal{R} = \{(x_1,\ldots,x_{n_X}) \in \{0,1\}^{n_X}, (y_1,\ldots,y_{n_Y}) \in \{0,1\}^{n_Y} : |t_{n_X,n_Y}| > F^{-1}_{\mathcal{N}(0,1)}(1-\alpha/2)\}.$$

• **Test exact de Fisher**:
Soit $(X_1,\ldots,X_{n_X})$ un échantillon i.i.d. de loi $\mathcal{B}(p_X)$ où $p_X \in]0,1[$. Soit $(Y_1,\ldots,Y_{n_Y})$ un échantillon i.i.d. de loi $\mathcal{B}(p_Y)$ où $p_Y \in]0,1[$. On souhaite tester au niveau α l'hypothèse nulle H_0: $p_X = p_Y$ contre l'hypothèse alternative H_1: $p_X \neq p_Y$.
Notons $n = n_X + n_Y$ et résumons les observations dans une table de contingence comme suit:

	échantillon 1	échantillon 2	total
succès (codé '1')	a	b	$a+b$
échec (codé '0')	c	d	$c+d$
total	$a+c$	$b+d$	$a+b+c+d=n$

Fisher a montré que, si les totaux marginaux sont fixés, si H_0 est vraie, la probabilité d'obtenir cette configuration est donnée par la loi hypergéometrique:

$$\frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}.$$

Les travaux de Fisher ont eu lieu avant ceux de Neyman et Pearson... et ne reposent pas sur la construction d'une région critique. Fisher calcule directement la p-valeur en la définissant comme la probabilité d'obtenir des résultats au moins autant en désaccord avec H_0 alors que H_0 est vraie. L'approche utilisée par la fonction `fisher.test` du logiciel `R` consiste à calculer la p-valeur en sommant les probabilités de toutes les tables ayant une probabilité inférieure ou égale à celle de la table observée.

• **Test d'homogénéité du χ^2 à J échantillons**:
D'une manière générale, 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 $j \in \{1, \ldots, J\}$ le vecteur des différents effectifs où $N_k^{(j)} = \sum_{i=1}^{n_j} I(X_i^{(j)} = a_k)$. Par définition, le vecteur $(N_1^{(j)}, \ldots, N_K^{(j)}$ suit une loi multinomiale de paramètres $(n_j, p_1^{(j)}, \ldots, p_K^{(j)})$ pour $j \in \{1, \ldots, J\}$. 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 $(p_1^{(j)}, \ldots, p_K^{(j)})$ sont identiques pour $j = 1, \ldots, J$. 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 $(p_1, \ldots, p_K)$ 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

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

La fonction `prop.test` (package `stats` chargé par défaut lors du lancement 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,
- les tailles n_X et n_Y des deux échantillons.