Université de Strasbourg
M1 - Magistère
Statistique - études de cas

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

Sujet 9: Tests de comparaison de deux moyennes à deux échantillons indépendants

• **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 deux moyennes à deux échantillons indépendants**:
Soient deux variables X et Y indépendantes dont on souhaite comparer les valeurs moyennes sur la base d'échantillons i.i.d. Formellement, on souhaite tester l'hypothèse nulle H_0: $\mathbb{E}[X] = \mathbb{E}[Y]$ contre l'hypothèse alternative H_1: $\mathbb{E}[X] \neq \mathbb{E}[Y]$ au niveau α.
Soit $(X_1, \ldots, X_{n_X})$ un échantillon i.i.d. distribué comme la variable X. Soit $(Y_1, \ldots, Y_{n_Y})$ un échantillon i.i.d. distribué comme la variable Y, indépendant de $(X_1, \ldots, X_{n_X})$. Soit

$$\overline{X}_{n_X} = \frac{1}{n_X} \sum_{i=1}^{n_X} X_i$$

un estimateur de $\mathbb{E}[X]$ et soit

$$\overline{Y}_{n_Y} = \frac{1}{n_Y}\sum_{i=1}^{n_Y} Y_i$$

un estimateur de $\mathbb{E}[Y]$. Notons

$$S'^2_{X,n_X} = \frac{1}{n_X - 1}\sum_{i=1}^{n_X}\left(\overline{X}_{n_X} - X_i\right)^2$$

un estimateur de $\mathrm{Var}(X)$ et

$$S'^2_{Y,n_Y} = \frac{1}{n_Y - 1}\sum_{i=1}^{n_Y}\left(\overline{Y}_{n_Y} - Y_i\right)^2$$

un estimateur de $\mathrm{Var}(Y)$. Notons également

$$S'^2_{n_X+n_Y} = \frac{(n_X-1)S'^2_{X,n_X} + (n_Y-1)S'^2_{Y,n_Y}}{n_X+n_Y-2} = \frac{1}{n_X+n_Y-2}\left(\sum_{i=1}^{n_X}\left(\overline{X}_{n_X} - X_i\right)^2 + \sum_{i=1}^{n_Y}\left(\overline{Y}_{n_Y} - Y_i\right)^2\right)$$

• **1er cas:** $n_X \geq 30$ **et** $n_Y \geq 30$:
Soit la statistique de test

$$T_{n_X,n_Y} = \frac{\overline{X}_{n_X} - \overline{Y}_{n_Y}}{\sqrt{\frac{S'^2_{X,n_X}}{n_X} + \frac{S'^2_{Y,n_Y}}{n_Y}}}.$$

La région critique est

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

en notant $F^{-1}_{\mathcal{N}(0,1)}(1-\alpha/2)$ le quantile d'ordre $(1-\alpha/2)$ de la loi gaussienne centrée réduite.

• **2ème cas:** $X \sim \mathcal{N}(\mathbb{E}[X],\sigma^2)$ **et** $Y \sim \mathcal{N}(\mathbb{E}[Y],\sigma^2)$**, autrement écrit les deux populations sont gaussiennes et de variances identiques (on parle alors d'homoscédasticité)**:
Soit la statistique de test

$$T_{n_X,n_Y} = \sqrt{\frac{n_X n_Y}{n_X+n_Y}}\,\frac{\overline{X}_{n_X} - \overline{Y}_{n_Y}}{\sqrt{S'^2_{n_X+n_Y}}}.$$

La région critique est

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

en notant $F^{-1}_{T(n_X+n_Y-2)}(1-\alpha/2)$ le quantile d'ordre $(1-\alpha/2)$ de la loi de Student à (n_X+n_Y-2) degrés de liberté.

• **3ème cas:** $X \sim \mathcal{N}(\mathbb{E}[X],\sigma_X^2)$ **et** $Y \sim \mathcal{N}(\mathbb{E}[Y],\sigma_Y^2)$**, autrement écrit les deux populations sont gaussiennes et de variances différentes (on parle alors d'hétéroscédasticité):**
La question de comparer les moyennes de deux échantillons gaussiens indépendants lorsque les variances diffèrent porte le nom de problème de Behrens-Fisher (d'après l'astronome Berhens

qui s'y est intéressé en 1929 et le statisticien Fisher qui s'y est intéressé en 1935). Il existe maintenant des solutions exactes à ce problème mais ici nous ne considérons que deux solutions approchées. Soit la statistique de test

$$T_{n_X,n_Y} = \frac{\overline{X}_{n_X} - \overline{Y}_{n_Y}}{\sqrt{\frac{S'^2_{X,n_X}}{n_X} + \frac{S'^2_{Y,n_Y}}{n_Y}}}.$$

La solution de Hsu-Scheffé consiste à prendre la région critique

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

en notant $F^{-1}_{T(\min(n_X,n_Y)-1)}(1-\alpha/2)$ le quantile d'ordre $(1-\alpha/2)$ de la loi de Student à $(\min(n_X, n_Y)-1)$ degrés de liberté.
La solution d'Aspin-Welch consiste à prendre la région critique

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

où ν est l'entier le plus proche de

$$\frac{\left(S'^2_{X,n_X}/n_X + S'^2_{Y,n_Y}/n_Y\right)^2}{\left(S'^2_{X,n_X}/n_X\right)^2 \times \frac{1}{n_X-1} + \left(S'^2_{Y,n_Y}/n_Y\right)^2 \times \frac{1}{n_Y-1}}$$

et en notant $F^{-1}_{T(\nu)}(1-\alpha/2)$ le quantile d'ordre $(1-\alpha/2)$ de la loi de Student à ν degrés de liberté.

• **Test de Wilcoxon-Mann-Whitney (ou test U de Mann-Whitney ou encore test de la somme des rangs de Wilcoxon)**:
Il s'agit à proprement parler d'un test de position. Soient deux variables X et Y indépendantes de fonction de répartition respective F et G que l'on suppose toutes deux continues. On suppose de plus que $G(x) = F(x - \delta)$ pour tout $x \in \mathbb{R}$. On souhaite comparer les positions respectives de X et Y sur la base d'échantillons i.i.d. Formellement, on souhaite tester l'hypothèse nulle H_0: $\delta = 0$ contre l'hypothèse alternative H_1: $\delta \neq 0$ au niveau α. Sous H_0, les deux échantillons à comparer sont donc issus de la même distribution.
Soit $(X_1, \ldots, X_{n_X})$ un échantillon i.i.d. distribué comme la variable X. Soit $(Y_1, \ldots, Y_{n_Y})$ un échantillon i.i.d. distribué comme la variable Y, indépendant de $(X_1, \ldots, X_{n_X})$. Ce test repose donc sur l'idée que, si H_0 est vraie, en mélangeant les valeurs obtenues dans les deux échantillons et en les ordonnant alors par valeurs croissantes, on doit obtenir un mélange homogène des deux échantillons. Pour $i \in \{1, \ldots, n_X\}$, notons R_{X_i} le rang de X_i dans l'échantillon aggloméré $(X_1, \ldots, n_Y, 1, \ldots, n_Y)$. De même, pour $j \in \{1, \ldots, n_Y\}$, notons R_{Y_j} le rang de Y_j dans l'échantillon aggloméré $(X_1, \ldots, n_Y, 1, \ldots, n_Y)$. Soit la statistique de test de la somme des rangs (proposée par Wilcoxon)

$$W_{n_X,n_Y} = \sum_{j=1}^{n_Y} R_{Y_j}.$$

Par des calculs directs (mais non triviaux), on peut montrer que $\mathbb{E}_{\delta=0}\left[W_{n_X,n_Y}\right] = \dfrac{n_Y(n_X + n_Y + 1)}{2}$ et que $\mathrm{Var}_{\delta=0}\left(W_{n_X,n_Y}\right) = \dfrac{n_X n_Y(n_X + n_Y + 1)}{12}$. La loi de W_{n_X,n_Y} sous H_0 ne dépend que de n_X et n_Y et est tabulée. Notons w_{n_X,n_Y} la réalisation de W_{n_X,n_Y}. La région critique est

$$\mathcal{R} = \{(x_1, \ldots, x_{n_X}, y_1, \ldots, y_{n_Y}) \in \mathbb{R}^{n_X+n_Y} : w_{n_X,n_Y} < k_{\alpha/2,n_X,n_Y} \text{ ou } w_{n_X,n_Y} > K_{\alpha/2,n_X,n_Y}\}$$

où $k_{\alpha/2,n_X,n_Y}$ et $K_{\alpha/2,n_X,n_Y}$ satisfont respectivement $\mathbb{P}_{\delta=0}(W_{n_X,n_Y} < k_{\alpha/2,n_X,n_Y}) = \alpha/2$ et $\mathbb{P}_{\delta=0}(W_{n_X,n_Y} > K_{\alpha/2,n_X,n_Y}) = \alpha/2$ et sont tabulés ou déterminés par ordinateur. De plus, dès lors que $n_X \geq 8$ et $n_Y \geq 8$, on peut approximer la loi de W_{n_X,n_Y} sous H_0 par la loi gaussienne.
NB: pour information, ce test est parfois présenté dans la version U de Mann-Whitney. La statistique de test U_{n_X,n_Y} compte le nombre de couples (X_i, Y_j) tels que $Y_j X_i$ a un rang supérieur à X_i:

$$U_{n_X,n_Y} = \sum_{i=1}^{n_X} \sum_{i=1}^{n_Y} I(R_{Y_j} > R_{X_i}) .$$

On peut montrer que

$$W_{n_X,n_Y} = U_{n_X,n_Y} + \frac{n_Y(n_Y+1)}{2}.$$

La fonction `wilcox.test` du logiciel `R` implémente les versions exacte et asymptotique de 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.

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

4. Tests non-paramétriques: faire varier la loi servant à générer les données ainsi que la valeur de son/ses paramètre(s).

Vous veillerez à faire varier tour à tour:

- le risque de 1ère espèce α,

- la taille de l'échantillon n.