Université de Strasbourg
M1 - Magistère
Statistique - études de cas

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

Sujet 10: tests de comparaison de 2 variances 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

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

• **Test de comparaison de 2 variances**:
Soient X et Y deux variables aléatoires indépendantes. On souhaite tester l'égalité de leurs variances respectives. Formellement, on souhaite tester l'hypothèse nulle H_0: $\mathrm{Var}(X) = \mathrm{Var}(Y)$ 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

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

un estimateur de $\mathrm{Var}(X)$ avec $\overline{X}_{n_X} = \dfrac{1}{n_X}\sum_{i=1}^{n_X} X_i$. Soit

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

un estimateur de $\mathrm{Var}(Y)$ avec $\overline{Y}_{n_Y} = \dfrac{1}{n_Y}\sum_{i=1}^{n_Y} Y_i$. Notons également

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

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

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

Sous H_0, la statique T_{n_X,n_Y} converge en loi vers une variable de loi $\mathcal{N}(0,1)$. Lorsque H_0 n'est pas satisfaite, la statistique T_{n_X,n_Y} diverge presque-sûrement vers ∞.

- Lorsque H_1: $\mathrm{Var}(X) \neq \mathrm{Var}(Y)$, la région critique est

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

- Lorsque H_1: $\mathrm{Var}(X) > \mathrm{Var}(Y)$, la région critique est

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

- Lorsque H_1: $\mathrm{Var}(X) < \mathrm{Var}(Y)$, la région critique est

$$\mathcal{R} = \left\{(x_1,\ldots,x_{n_X}) \in \mathbb{R}^{n_X}, (y_1,\ldots,y_{n_Y}) \in \mathbb{R}^{n_Y} : t_{n_X,n_Y} < F^{-1}_{\mathcal{N}(0,1)}(\alpha)\right\}.$$

2ème cas (test de Fisher ou de Fisher-Snedecor): $X \sim \mathcal{N}(\mathbb{E}[X], \mathrm{Var}(X))$ et $Y \sim \mathcal{N}(\mathbb{E}[Y], \mathrm{Var}(Y))$.
Soit la statistique de test:

$$T_{n_X,n_Y} = \frac{S'^2_{n_X}}{S'^2_{n_Y}}.$$

Sous H_0, la statique T_{n_X,n_Y} suit la loi $F(n_X,n_Y)$ dite de Fisher à n_X et n_Y degrés de liberté. Heuristiquement, lorsque H_0 est vraie, le rapport T_{n_X,n_Y} ne doit pas trop s'éloigner de 1.

- Lorsque H_1: $\mathrm{Var}(X) \neq \mathrm{Var}(Y)$, la région critique est

$$\mathcal{R} = \left\{(x_1,\ldots,x_{n_X}) \in \mathbb{R}^{n_X}, (y_1,\ldots,y_{n_Y}) \in \mathbb{R}^{n_Y} : t_{n_X,n_Y} > F^{-1}_{F(n_X,n_Y)}(1-\alpha/2) \text{ ou } t_{n_X,n_Y} < F^{-1}_{F(n_X,n_Y)}(\alpha/2)\right\}$$

- Lorsque H_1: $\mathrm{Var}(X) > \mathrm{Var}(Y)$, la région critique est

$$\mathcal{R} = \left\{(x_1,\ldots,x_{n_X}) \in \mathbb{R}^{n_X}, (y_1,\ldots,y_{n_Y}) \in \mathbb{R}^{n_Y} : t_{n_X,n_Y} > F^{-1}_{F(n_X,n_Y)}(1-\alpha)\right\}.$$

- Lorsque H_1: $\mathrm{Var}(X) < \mathrm{Var}(Y)$, la région critique est

$$\mathcal{R} = \left\{(x_1, \ldots, x_{n_X}) \in \mathbb{R}^{n_X}, (y_1, \ldots, y_{n_Y}) \in \mathbb{R}^{n_Y} : t_{n_X, n_Y} < F^{-1}_{F(n_X, n_Y)}(\alpha)\right\}.$$

• **Test de comparaison de K variances**:
Soient $X^{(k)}$, pour $k = 1, \ldots, K$, des variables aléatoires indépendantes où $X^{(k)} \sim \mathcal{N}(\mathbb{E}[X^{(k)}], \mathrm{Var}(X^{(k)}))$, pour $k = 1, \ldots, K$. On souhaite tester l'égalité de leurs variances respectives. Formellement, on souhaite tester au niveau α l'hypothèse nulle H_0: $\mathrm{Var}\left(X^{(k)}\right) = \sigma^2$ pour $k = 1, \ldots, K$ contre l'hypothèse alternative H_1: $\exists k_1 \neq k_2 \in \{1, \ldots, K\}$ tels que $\mathrm{Var}\left(X^{(k_1)}\right) \neq \mathrm{Var}\left(X^{(k_2)}\right)$. Pour $k = 1, \ldots, K$, soit $(X_1^{(k)}, \ldots, X_{n_k}^{(k)})$ un échantillon i.i.d. distribué comme la variable $X^{(k)}$. Pour $k = 1, \ldots, K$, soit

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

un estimateur de $\mathrm{Var}\left(X^{(k)}\right)$ avec

$$\overline{X}^{(k)}_{n_k} = \frac{1}{n_k} \sum_{i=1}^{n_k} X_i^{(k)}.$$

1er cas: Test de Bartlett.
Soit la statistique de test:

$$T_{n_1,\ldots,n_K} = \frac{(n-K)\left[\log\left(\frac{1}{n-K}\sum_{k=1}^K (n_k - 1) S'^2_{k,n_k}\right) - \frac{1}{n-K}\sum_{k=1}^K (n_k - 1)\log\left(S'^2_{k,n_k}\right)\right]}{1 + \frac{1}{3(K-1)}\left(\sum_{k=1}^K \frac{1}{n_K - 1} - \frac{1}{n-K}\right)}$$

en notant $n = \sum_{k=1}^K n_k$. Sous H_0, la statistique $T_{n_1,\ldots,n_K}$ suit la loi $\chi^2(K-1)$. Notons $t_{n_1,\ldots,n_K}$ la réalisation de $T_{n_1,\ldots,n_K}$. La région critique est:

$$\mathcal{R} = \left\{(x_1, \ldots, x_{n_k})_{k=1,\ldots,K} : t_{n_1,\ldots,n_K} > F^{-1}_{\chi^2(K-1)}(1-\alpha)\right\}.$$

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

2ème cas: Test de Levene.
Introduisons:

$$\begin{aligned} Z_i^{(k)} &= \left|X_i^{(k)} - \overline{X}^{(k)}_{n_k}\right| \\ \overline{Z}^{(k)}_{n_k} &= \frac{1}{n_k}\sum_{i=1}^{n_k} Z_i^{(k)} \\ \overline{Z}_n &= \frac{1}{n}\sum_{k=1}^K \sum_{i=1}^{n_k} Z_i^{(k)} = \frac{1}{n}\sum_{k=1}^K n_k \overline{Z}^{(k)}_{n_k}. \end{aligned}$$

Soit la statistique de test:

$$T_{n_1,\ldots,n_K} = \frac{n-K}{K-1} \frac{\sum_{k=1}^K \left(\overline{Z}^{(k)}_{n_k} - \overline{Z}_n\right)^2}{\sum_{k=1}^K \sum_{i=1}^{n_k} \left(Z_i^{(k)} - \overline{Z}^{(k)}_{n_k}\right)^2}.$$

Sous H_0, la statistique $T_{n_1,\ldots,n_K}$ suit la loi de Fisher $F(K-1, n-K)$. Notons $t_{n_1,\ldots,n_K}$ la réalisation de $T_{n_1,\ldots,n_K}$. La région critique est:

$$\mathcal{R} = \left\{ (x_1, \ldots, x_{n_k})_{k=1,\ldots,K} : t_{n_1,\ldots,n_K} > F^{-1}_{F(K-1,n-K)}(1-\alpha) \right\}.$$

La fonction `leveneTest` du package `car` du logiciel `R` implémente ce test ainsi que la fonction `levene.test` du package `lawstat`.

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,
- la variabilité du phénomène.

Le test de Levene a la réputation d'être robuste aux écarts à la normalité. Qu'en pensez-vous?