Tests sur deux échantillons : comparer moyennes et variances

Frédérique Leblanc&Victor Léger

Les données :

- X est observé pour n_X individus :

$$x_1, x_2, \ldots, x_{n_X}$$

- Y est observé pour n_Y individus :

$$y_1, y_2, \ldots, y_{n_Y}$$

Deux situations possibles :

- mêmes individus observés en X et Y : **Echantillons appariés**
- différents individus observés en X et Y : **Echantillons indépendants**

1. Comparer les moyennes de X et Y (cas appariés et indép.)
2. Comparer les variances de X et Y (cas indép.)
3. Comparer les distributions de X et Y (cas indép. voir chap suivant-non traité cette année)

Exemples : dans le fichier de données du poly :
1) poids à 15 ans et poids à 20 ans $==>$ appariés.
question possible : le poids à 15 ans est-il en moyenne le même qu'à 20 ans ?
2) poids chez les filles et poids chez les garçons $==>$ indépendants.
question possible : le sexe a t-il en moyenne un effet sur le poids ?

On notera $E(X) = \mu_X$, $E(Y) = \mu_Y$, $V(X) = \sigma_X^2$ et $V(Y) = \sigma_Y^2$.
Selon le cas on posera les modèles suivants :

1. **Appariés** : Soit $D = X - Y$ on suppose D de loi $\mathcal{N}(\mu_D, \sigma_D^2)$ où $\mu_D = \mu_X - \mu_Y$ et σ_D sont inconnus.
2. **Indépendants** :
   i) X et Y sont supposées de loi resp. $\mathcal{N}(\mu_X, \sigma_X^2)$ et $\mathcal{N}(\mu_Y, \sigma_Y^2)$
   ii) X et Y indépendantes
   iii) une des conditions suivantes
   - σ_X et σ_Y connus
   - $\sigma_X = \sigma_Y$ si σ_X et σ_Y inconnus et petits échantillons ($n < 30$)
   - σ_X et σ_Y inconnus et grands échantillons (au moins de tailles 100)

On veut tester

$$\mathcal{H}_0 : \mu_D \leq 0 \qquad \mathcal{H}_1 : \mu_D > 0 \qquad W_\alpha = \{T > t_{n-1,1-\alpha}\}$$

$$\mathcal{H}_0 : \mu_D \geq 0 \qquad \mathcal{H}_1 : \mu_D < 0 \qquad W_\alpha = \{T < -t_{n-1,1-\alpha}\}$$

$$\mathcal{H}_0 : \mu_D = 0 \qquad \mathcal{H}_1 : \mu_D \neq 0 \qquad W_\alpha = \{|T| > t_{n-1,1-\alpha/2}\}$$

On applique les résultats des test sur un seul échantillon à l'échantillon des différences : $d_1, d_2, ..., d_n$.
Statistique de test : $T = \bar{D}\sqrt{n}/S'_D$
Loi de T sous $\mathcal{H}_0$: Student à $n-1$ d.l.

Modèle :
X de loi $\mathcal{N}(\mu_X, \sigma_X^2)$
Y de loi $\mathcal{N}(\mu_Y, \sigma_Y^2)$
X et Y indépendantes

$$\mathcal{H}_0 : \sigma_X^2 = \sigma_Y^2 \qquad \mathcal{H}_1 : \sigma_X > \sigma_Y \qquad W_\alpha = \{T > f_{n_X-1, n_Y-1, 1-\alpha}\}$$

$$\mathcal{H}_0 : \sigma_X = \sigma_Y \qquad \mathcal{H}_1 : \sigma_X < \sigma_Y \qquad W_\alpha = \{T < f_{n_X-1, n_Y-1, \alpha}\}$$

$$\mathcal{H}_0 : \sigma_X = \sigma_Y \qquad \mathcal{H}_1 : \sigma_X \neq \sigma_Y$$

$$W_\alpha = \{T > f_{n_X-1, n_Y-1, 1-\alpha/2} \;\; ou \;\; T < f_{n_X-1, n_Y-1, \alpha/2}\}$$

Statistique de Test : $T = S_X'^2 / S_Y'^2$

loi de T sous $\mathcal{H}_0$: loi de Fisher de paramètres $(n_X - 1, n_Y - 1)$ notée $\mathcal{F}_{n_X-1,n_Y-1}$

Régions de rejet :

$$\{T > f_{n_X-1,n_Y-1,1-\alpha}\} \ , \ \{T < f_{n_X-1,+n_Y-1,\alpha}\}$$

$$\{T > f_{n_X-1,n_Y-1,1-\alpha/2}\} \cup \{T < f_{n_X-1,+n_Y-1,\alpha/2}\}$$

Modèle :
X de loi $\mathcal{N}(\mu_X, \sigma_X^2)$ et Y de loi $\mathcal{N}(\mu_Y, \sigma_Y^2)$
X et Y indépendantes
et une des conditions iii) satisfaite

Tests de comparaison de moyennes :

$$\mathcal{H}_0 : \mu_X - \mu_Y = 0 \qquad \mathcal{H}_1 : \mu_X - \mu_Y > 0$$

$$\mathcal{H}_0 : \mu_X - \mu_Y = 0 \qquad \mathcal{H}_1 : \mu_X - \mu_Y < 0$$

$$\mathcal{H}_0 : \mu_X - \mu_Y = 0 \qquad \mathcal{H}_1 : \mu_X - \mu_Y \neq 0$$

Estimateur de $\mu_X - \mu_Y$: $\bar{X} - \bar{Y}$
Régions de Rejet : $\{T > ...\}, \{T < ...\}$ et $\{|T| > ..\}$
Ensuite plusieurs situations :

- les variances sont connues (situation rare)
- elles sont inconnues mais supposées égales (petits échantillons possibles)
- elles sont inconnues et non nécessairement égales (mais avec grands échantillons $n > 100$)

Statistique de Test : Elle dépend de la situation et est l'estimateur centré et réduit sous $\mathcal{H}_0$

loi de $\bar{X} - \bar{Y}$:

$$\mathcal{N}\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right)$$

loi de $\bar{X} - \bar{Y}$ **sous** $\mathcal{H}_0$:

$$\mathcal{N}\left(0, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right)$$

loi de $\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}}$ **sous** $\mathcal{H}_0$:

$$\mathcal{N}(0, 1)$$

1. **variances connues :**

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$$

Loi de T sous $\mathcal{H}_0$: $\mathcal{N}(0,1)$

2. **variances inconnues et grands échantillons :**

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}}$$

Loi de T sous $\mathcal{H}_0$: approximativement $\mathcal{N}(0,1)$

3. **variances inconnues et égales :**

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\Sigma^2(n_X^{-1} + n_Y^{-1})}} \qquad \text{avec} \qquad \Sigma^2 = \frac{n_X S_X^2 + n_Y S_Y^2}{n_X + n_Y - 2}$$

Loi de T sous $\mathcal{H}_0$: Student $\mathcal{T}_{n_X+n_Y-2}$

Pour les situations 1 et 2 les régions de rejet sont

$$\{T > u_{1-\alpha}\} , \quad \{T < -u_{1-\alpha}\} , \quad \{|T| > u_{1-\alpha/2}\}$$

et les p-valeurs

$$1 - \phi(T_{calc}) , \quad \phi(T_{calc}) , \quad 2 - 2\Phi(|T_{calc}|)$$

Pour la situation 3 les régions de rejet sont

$$\{T > t_{n_X+n_Y-2,1-\alpha}\} , \{T < -t_{n_X+n_Y-2,1-\alpha}\} , \{|T| > t_{n_X+n_Y-2,1-\alpha/2}\}$$

et les p-valeurs

$$1 - F(T_{calc}) , \quad F(T_{calc}) , \quad 2 - 2F(|T_{calc}|)$$

où F est la FdR de la $\mathcal{T}_{n_X+n_Y-2}$

Exercice :

I) La taille d'une fille est elle plus variable que celle d'un garçon ?
II) Est-elle en moyenne différente chez les filles et les garçons ?

Avec les données suivantes (voir poly) on a pris parmi les étudiants ayant répondu à l'enquête, un échantillon de 14 filles et 6 garçons :

chez les filles : $n_X = 14 \quad , \sum x_i = 2288, \sum x_i^2 = 376226$
chez les garçons : $n_Y = 6 \quad , \sum y_i = 1145, \sum y_i^2 = 219231$

Comme les échantillons sont de petites tailles, pour comparer **les moyennes théoriques** μ_X et μ_Y avec un test statistique et répondre à la question II) il faut d'abord se donner les hypothèses de modélisation :

- i) X la taille d'une femme est supposée aléatoire et de loi $\mathcal{N}(\mu_X, \sigma_X^2)$ et celle d'un homme Y de loi $\mathcal{N}(\mu_Y, \sigma_Y^2)$ (μ_X, μ_Y, σ_X et σ_Y sont inconnus)
- ii) X et Y sont indépendantes
- ii) $\sigma_X = \sigma_Y$

Test de comparaison des moyennes pour répondre à II)

$$\mathcal{H}_0 : \mu_X - \mu_Y = 0 \qquad \mathcal{H}_1 : \mu_X - \mu_Y \neq 0$$

L'alternative $\mathcal{H}_1$ testée ici traduit "il y a en moyenne un effet du sexe sur la taille d'une personne" tandis que l'hypothèse nulle traduit au contraire "pas d'effet en moyenne du sexe sur la taille".
Région de Rejet au seuil α : $W_\alpha = \{T > t_{n_X+n_Y-2}\}$ avec

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\Sigma^2(n_X^{-1} + n_Y^{-1})}} \qquad \text{avec} \qquad \Sigma^2 = \frac{n_X S_X^2 + n_Y S_Y^2}{n_X + n_Y - 2}$$

et la p-valeur du test est donnée par $\alpha^* = 2(1 - F(T_{calc}))$ (voir diapo; 11)

Les calculs pour les deux échantillons observés

$$\bar{x} = 2288/14 = 163.4; \quad s_x^2 = 164.39; \quad s_x'^2 = 177.03; \quad s_x' = 13.3$$

$$\bar{y} = 1145/6 = 190.8; \quad s_y^2 = 121.14; \quad s_y'^2 = 145.37; \quad s_y' = 12.0$$

$$\Sigma_{calc}^2 = \frac{14 \cdot 164.39 + 6 \cdot 121.14}{14 + 6 - 2} = 178.07$$

et

$$T_{calc} = \frac{163.4 - 190.08}{\sqrt{178.07 \cdot (1/14 + 1/6)}} = -4.2$$

et la p-valeur du test : $\alpha^* = 2(1 - F(|T_{calc}|)) = 5 \cdot 10^{-4}$ s'obtient avec la calculette ou R : `2-2*pt(abs(tcalc),18)`
On conclut donc qu'en moyenne le sexe a un effet sur la taille avec un risque de se tromper α dès que $\alpha > 5 \cdot 10^{-4}$ (soit de façon statistiquement très significative)

Pour comparer les moyennes nous avons supposé $\sigma_X = \sigma_Y$: est-ce une hypothèse de modélisation raisonnable ?
Pour répondre on fait **un test bilatéral de comparaisons de variances :**

$$\mathcal{H}_0 : \sigma_X^2 = \sigma_Y^2 \qquad \qquad \mathcal{H}_1 : \sigma_X^2 \neq \sigma_Y^2$$

$$W_\alpha = \{T > f_{n_X-1,n_Y-1,1-\alpha/2} \ \ ou \ \ T < f_{n_X-1,n_Y-1,\alpha/2}\}$$

A.N. : $T_{calc} = s_x'^2/s_y'^2 = 177.03/145.37 = 1.22$, on chosit de tester avec $\alpha = 0.05$ et on lit $f_{13,5,1-0.05/2} = f_{13,5,0.975} = 6.49$ on calcule $f_{13,5,0.05/2} = 1/f_{5,13,1-0.05/2} = 1/3.77 = 0.27$.
Comme $T_{calc} = 177.03 \notin W_{5\%}$ *on ne rejette pas* $\sigma_X = \sigma_Y$ *dans un test où le risque de rejet à tort vaut* 5%. *Supposer* $\sigma_X = \sigma_Y$ *est donc une hyp. de modélisation raisonnable.*

Pour répondre à la question I) on fait un
Test unilatéral sur les variances :

$$\mathcal{H}_0 : \sigma_X^2 = \sigma_Y^2 \qquad \qquad \mathcal{H}_1 : \sigma_X^2 > \sigma_Y^2$$

$$W_\alpha = \{T > f_{n_X-1,n_Y-1,1-\alpha}\}$$

avec une p-valeur α^* :

$$\alpha^* = 1 - F_{n_X-1,n_Y-1}(T_{calc}) = \texttt{1-pf(177.03,13,5)} = 44.8\%$$

. Donc à moins de conclure à tort avec un risque d'au moins 44.8% on ne peut montrer que la taille d'une femme est plus variable que celle d'un homme.

A la question I) on répondra donc que rien de statistiquement significatif ne permet de dire que la taille d'une femme est plus variable que celle d'un homme.