Tests sur deux échantillons : comparer moyennes et variances

Frédérique Leblanc

Les données :

- X est observé pour n_X individus :

$$x_1, x_2, ..., x_{n_X}$$

- Y est observé pour n_Y individus :

$$y_1, y_2, ..., y_{n_Y}$$

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)

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) $\sigma_X = \sigma_Y$ si inconnus et petits échantillons ($n < 30$)

On veut tester

$$\mathcal{H}_0 : \mu_D \leq 0 \qquad \mathcal{H}_1 : \mu_D > 0$$

$$\mathcal{H}_0 : \mu_D \geq 0 \qquad \mathcal{H}_1 : \mu_D < 0$$

$$\mathcal{H}_0 : \mu_D = 0 \qquad \mathcal{H}_1 : \mu_D \neq 0$$

On applique les résultats des test à 1 échantillon sur 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 $\mathcal{N}(\mu_Y, \sigma_Y^2)$

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

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

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

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

loi de T sous $\mathcal{H}_0$: loi de Fisher de para. $\mathcal{F}_{n_X-1,n_Y-1}$

Regions 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 $\mathcal{N}(\mu_X, \sigma_X^2)$
Y de loi $\mathcal{N}(\mu_Y, \sigma_Y^2)$
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 (grands echantillons $n > 100$)

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 $\mathcal{H}_O$: $\mathcal{N}(0,1)$

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 $\mathcal{H}_O$: 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 $\mathcal{H}_O$: 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 :

La taille d'une fille est elle plus variable que celle d'un garçon ?
Est-elle différente chez les filles et les garçons ?
Avec les données suivantes (voir poly) :
chez les filles : $n_X = 14 \quad , \sum x_i = 2288, \sum x_i^2 = 376226$
chez les garçons : $n_Y = 6 \quad , \sum y_i = 1145, \sum y_i^2 = 219231$