# Points importants du cours 3

Vincent Guillemot

22 janvier 2009

## 1 Tests d'hypothèses

### 1.1 Exemple préliminaire

(cf. cours précédent)

### 1.2 Démarche générale

La démarche générale à suivre se décline en 4 étapes :

1. Choix de $\mathcal{H}_0$ et de $\mathcal{H}_1$
2. Choix de la variable aléatoire répondant à la question posée. On parle de **statistique** (de test).
3. Définition de la région critique (*i.e.* de rejet de $\mathcal{H}_0$) en fonction du risque $\alpha$ et de $\mathcal{H}_1$.
4. Décision.

Ce cadre sera respecté pour les tests données en exemple dans la suite. Nous verrons deux grands types de tests :

- les tests paramétriques formulant une hypothèse supplémentaire sur la nature statistique de la(es) variable(s) aléatoire(s) étudiée(s),
- les tests non paramétriques qui permettent de prendre une décision quelle que soit la distribution de cette(es) variable(s).

### 1.3 Test paramétrique

Un test paramétrique très utilisé dans les études de données biologiques est le test de Student, car il permet de répondre à une question souvent formulée : "l'échantillon 1 est-il en moyenne significativement différent de l'échantillon 2" ? Ce test nécessite de formuler l'hypothèse de travail suivante sur les données (que nous accompagnons des notations nécessaires) : soit $X$ et $Y$ deux v.a.r. indépendantes telles que $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ et $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. On se donne de plus deux échantillons aléatoires de tailles respectives $n_X$ et $n_Y$ relatifs à $X$ et $Y$ : $(X_i)_{i=1,\dots,n_X}$ et $(Y_i)_{i=1,\dots,n_Y}$. Ce test se décline en plusieurs cas, dont deux nécessitent

de contrôler avec un premier test si les variances de deux échantillons sont significativement différentes.

**Attention :** La notation $\mathcal{RH}_0$ signifie que l'on rejette $\mathcal{H}_0$.

### 1.3.1 Comparaison de la variance de deux échantillons

Cas 1 : les moyennes $\mu_X$ et $\mu_Y$ sont connues.

1. $\mathcal{H}_0$ : "$\sigma_X^2 = \sigma_Y^2$" ; $\mathcal{H}_1$ : "$\sigma_X^2 \neq \sigma_Y^2$".
2. La statistique utilisée sera la suivante :
$$F_1 = \frac{S_{\mu_X}^2}{S_{\mu_Y}^2} \sim \mathcal{F}(n_X, n_Y)$$
3. La décision se formule donc ainsi au risque $\alpha$ :
$$F_1 \geq f_{1-\alpha/2}(n_X, n_Y) \Rightarrow \mathcal{RH}_0.$$

Cas 2 : les moyennes $\mu_X$ et $\mu_Y$ sont inconnues.

1. $\mathcal{H}_0$ : "$\sigma_X^2 = \sigma_Y^2$" ; $\mathcal{H}_1$ : "$\sigma_X^2 \neq \sigma_Y^2$".
2. La statistique utilisée sera la suivante :
$$F_2 = \frac{S_X^2}{S_Y^2} \sim \mathcal{F}(n_X - 1, n_Y - 1)$$
3. La décision se formule donc ainsi au risque $\alpha$ :
$$F_2 \geq f_{1-\alpha/2}(n_X - 1, n_Y - 1) \Rightarrow \mathcal{RH}_0.$$

### 1.3.2 Comparaison de la moyenne de deux échantillons

Cas 0 : les échantillons sont appariés.

**Définition 1 - *Échantillons appariés***

Deux échantillons sont dits appariés dès que les individus présents dans l'un et l'autre échantillon sont les mêmes. L'intérêt est principalement de mesurer un écart entre deux situations différentes.

1. $\mathcal{H}_0$ : "$\mu_D = 0$" ; $\mathcal{H}_1$ : "$\mu_D \neq 0$".
2. La statistique utilisée sera la suivante :
$$T_0 = \frac{\overline{D}}{\sqrt{\frac{s_D^2}{n-1}}} \sim \mathcal{T}(n-1)$$
3. La décision se formule donc ainsi au risque $\alpha$ :
$$T_0 \geq t_{1-\alpha/2}(n-1) \Rightarrow \mathcal{RH}_0.$$

Cas 1 : les variances sont connues et égales.

**Proposition 1 - *Combinaison linéaire de variables gaussiennes.***

Soit $X_1, ..., X_n$ $n$ variables aléatoires gaussiennes
- mutuellement indépendantes,
- et telles que $\forall i, X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$.

On choisit $n$ réels $a_1, ..., a_n$. Alors la variable $Y = \sum_{i=1}^{n} a_i X_i$ est une variable aléatoire gaussienne
- de moyenne $a_1\mu_1 + ... + a_n\mu_n$,
- et de variance $a_1^2\sigma_1^2 + ... + a_n^2\sigma_n^2$.

**Par exemple :** La variable $\overline{X} - \overline{Y}$ est une variable gaussienne :

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

1. $\mathcal{H}_0$ : "$\mu_X = \mu_Y$" ; $\mathcal{H}_1$ : "$\mu_X \neq \mu_Y$".
2. La statistique utilisée sera la suivante :

$$Z_1 = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim \mathcal{N}(0,1)$$

3. La décision se formule donc ainsi au risque $\alpha$ :

$$Z_1 \geq z_{1-\alpha/2} \Rightarrow \mathcal{RH}_0.$$

Cas 2 : les variances sont connues et différentes.

1. $\mathcal{H}_0$ : "$\mu_X = \mu_Y$" ; $\mathcal{H}_1$ : "$\mu_X \neq \mu_Y$".
2. La statistique utilisée sera la même :

$$Z_2 = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim \mathcal{N}(0,1)$$

3. La décision se formule donc ainsi au risque $\alpha$ :

$$Z_2 \geq z_{1-\alpha/2} \Rightarrow \mathcal{RH}_0.$$

**Remarque :** Quand $\sigma_X = \sigma_Y = \sigma$, attention à ne pas se laisser emporter !

Cas 3 : les variances sont inconnues et égales.

1. $\mathcal{H}_0$ : "$\mu_X = \mu_Y$" ; $\mathcal{H}_1$ : "$\mu_X \neq \mu_Y$".
2. La statistique utilisée sera la suivante :

$$T_1 = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{(n_X-1)S_X^2 + (n_Y-1)S_Y^2}{n_X+n_Y-2}\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} \sim \mathcal{T}(n_X + n_Y - 2)$$

3. La décision se formule donc ainsi au risque $\alpha$ :

$$T_1 \geq t_{1-\alpha/2}(n_X + n_Y - 2) \Rightarrow \mathcal{R}\mathcal{H}_0.$$

Cas 4 : les variances sont inconnues et différentes.

On utilise dans ce cas toujours une statistique qui va suivre une loi de Student, mais la version du test utile dans ce cas s'appelle le ***test d'Aspin-Welch***.

1. $\mathcal{H}_0$ : "$\mu_X = \mu_Y$" ; $\mathcal{H}_1$ : "$\mu_X \neq \mu_Y$".
2. La statistique utilisée sera la suivante :

$$T_2 = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}} \sim \mathcal{T}(m),$$

où le nombre de degrés de liberté $m$ est l'entier se rapprochant le plus de la quantité suivante :

$$m \approx \frac{c^2}{n_X - 1} + \frac{1 - c^2}{n_Y - 1}, \text{ avec } c = \frac{S_X^2/n_X}{S_X^2/n_X + S_Y^2/n_Y}.$$

3. La décision se formule donc ainsi au risque $\alpha$ :

$$T_2 \geq t_{1-\alpha/2}(m) \Rightarrow \mathcal{R}\mathcal{H}_0.$$