# TP4 : test multiple par rééchantillonnage

*ENSAE 3ème année, 2017*

*Applications du bootstrap et autres techniques de rééchantillonnage (cours de M. Roquain)*

Dans ce TP, nous allons traiter des données réelles (assez massives). On observe le niveau d'expression de $m = 6033$ gènes entre deux groupes d'individus : le groupe 0, de taille $n_0 = 50$, correspond à des individus "sains" ; le groupe groupe 1, de taille $n_1 = 52$, correspond à des individus "malades" (cancer de la prostate).

## Exercice 0 (Création de votre fichier)

Créer votre fichier `nom_prenom_TP4.R`. N'oubliez pas d'effectuer des commentaires dans votre fichier (que vous utilisiez un notebook ou pas).

## Exercice 1 (Chargement, représentation, modélisation des données)

1) Récupérer le fichier `prostate.rda` disponible sur ma page http://etienne.roquain.free.fr/teaching.html. Charger ensuite les données de la manière suivante :

```
load("prostate.rda")
X=prostate.x
n=dim(X)[1]
m=dim(X)[2]
n0=sum(prostate.y==0) # groupe des individus sains
n1=n-n0 # groupe des individus malades
```

2) Pour certains gènes, représenter leurs expressions à travers les deux groupes d'individus. A vue de nez, le niveau d'expression est-il différent entre les deux groupes pour certains gènes ? On pourra essayer le gène numéro `2619`.

Dans toute la suite, on modélisera les données avec le modèle vu en cours, sous l'hypothèse gaussienne multivariée. Plus spécifiquement, on suppose que

$$\mathcal{X}_n = (X_1, \dots, X_n) = (Y_1, \dots, Y_{n_0}, Z_1, \dots, Z_{n_1})$$

est un vecteur de variables dans $\mathbb{R}^m$ qui sont mutuellement indépendantes, que l'échantillon $(Y_1, \dots, Y_{n_0})$ est i.i.d. avec $Y_1 - \theta_1 \sim \mathcal{N}(0, \Gamma)$ et que l'échantillon $(Z_1, \dots, Z_{n_1})$ est i.i.d. avec $Z_1 - \theta_2 \sim \mathcal{N}(0, \Gamma)$, pour une certaine matrice de covariance $\Gamma$ inconnue. Le paramètre $\theta_0 = (\theta_{0,j})_{1 \le j \le m}$ (resp. $\theta_1 = (\theta_{1,j})_{1 \le j \le m}$) désigne le niveau d'expression théorique du groupe sain (resp. malade) pour tous les gènes.

On cherche ainsi à déterminer l'ensemble des gènes différentiellement exprimé, ou, de manière équivalente, son complémentaire

$$\mathcal{H}_0 = \{j \in \{1, \dots, m\} \ : \ \theta_{1,j} = \theta_{0,j}\}.$$

La méthode employée dans ce TP est celle des tests multiples.

## Exercice 2 (Test de student)

1) Pour tout $j \in \{1, \dots, m\}$, calculer la statistique de test de Student pour tester $H_{0,j}$ : "$\theta_{1,j} = \theta_{0,j}$" contre $H_{1,j}$ : "$\theta_{1,j} \neq \theta_{0,j}$". On pourra utiliser la fonction `t.test` avec la valeur `$stat` et vérifier "à la main" que le logiciel rend bien la quantité

$$T_n(\mathcal{X}_{n,j}) = (1/n_0 + 1/n_1)^{-1/2} \frac{\bar{Y}_j - \bar{Z}_j}{\left((n-2)^{-1}\left(\sum_{i=1}^{n_0}(Y_{i,j} - \bar{Y}_j)^2 + \sum_{i=1}^{n_1}(Z_{i,j} - \bar{Z}_j)^2\right)\right)^{1/2}},$$

où $\bar{Y}_j = n_0^{-1}\sum_{i=1}^{n_0} Y_{i,j}$ et $\bar{Z}_j = n_1^{-1}\sum_{i=1}^{n_1} Z_{i,j}$.

2) Tracer les $T_n(\mathcal{X}_{n,j}), 1 \leq j \leq m$, ainsi que les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ d'une loi de Student de paramètre $n-2$ pour $\alpha = 5\%$. Quelles sont les hypothèses $H_{0,j}$ rejetées au niveau 5% ?

3) Refaire 1) et 2) à partir du même jeu de données avec les lignes préalablement permutées au hasard. Que remarque t-on ?

## Exercice 3 (Procédure de Bonferroni)

1) Refaire le 2) et 3) de l'exercice 2 en appliquant cette fois un seuillage à l'aide des quantiles d'ordre $\alpha/(2m)$ et $1-\alpha/(2m)$ d'une loi de Student de paramètre $n-2$ pour $\alpha = 5\%$.

2) Donner la liste des gènes différentiellement exprimés selon la méthode de Bonferroni, ainsi que la garantie que l'on peut avoir sur cette liste en terme de faux positifs.

## Exercice 4 (Procédure de Romano-Wolf)

On rappelle que le seuil de Romano-Wolf est défini par le $(1-\alpha)$-quantile de la loi

$$\mathcal{L}\left(\sup_{1\leq j\leq m} |T_n(\mathcal{X}^\sigma_{n,j})| \,\middle|\, \mathcal{X}_n\right)$$

où $\mathcal{X}^\sigma_n = (X_{\sigma(1)}, \dots, X_{\sigma(n)})$ et $\sigma$ est une permutation aléatoire, uniformément distribuée dans les permutations de $\{1, \dots, n\}$.

1) Approcher le seuil de Romano-Wolf par la méthode de Monte-Carlo ($B = 100$).

2) Donner la liste des gènes différentiellement exprimés selon la méthode de Romano-Wolf. Comparer la liste obtenue à celle obtenue à l'exercice 3. Commenter.