# TP4 : Test par permutation dans le modèle cas/contrôle

*M2 Statistique, 2017-2018*

*Statistique mathématique en grande dimension et applications, cours de M. Roquain*

## Exercice 1 : Test par permutation

Soit $n_0 = 50$, $n_1 = 52$ et $n = n_0 + n_1$ et $Z \sim P_0$. On observe un échantillon

$$X = (X_1, \dots, X_n) = (X_1, \dots, X_{n_0}, X_{n_0+1}, \dots, X_n),$$

avec $(X_1, \dots, X_{n_0})$ i.i.d. de même loi que $Z$ et $(X_{n_0+1}, \dots, X_n)$ i.i.d. de même loi que $\delta + Z$. On cherche à tester $H_0 : \delta = 0$ contre $H_1 : \delta > 0$.

On considère la statistique de test

$$T(X) = \frac{n_1^{-1} \sum_{i=n_0+1}^{n} X_i - n_0^{-1} \sum_{i=1}^{n_0} X_i}{(1/n_1 + 1/n_0)^{1/2}}$$

1) Quelle est la loi de $T(X)$ sous $H_0$ lorsque $P_0 = \mathcal{N}(0,1)$ ?

2) Construire la fonction $T : x \mapsto T(x)$. Simuler une matrice $n \times B$ ($B = 1000$) avec des entrées $\mathcal{N}(0,1)$ et appliquer la fonction $T$ aux colonnes. Tracer l'histogramme des $B$ valeurs obtenues.

3) Simuler $X$ sous $H_0$ lorsque $P_0 = \mathcal{N}(0,1)$ comme suit.

```
X=rnorm(n)
```

Que fait le programme suivant ?

```
B=1000
Sigma=matrix(sapply(1:B,function(b) sample(n)),n,B,byrow=FALSE)
statperm=sapply(1:B,function(b) T(X[Sigma[,b]]))
hist(statperm,freq=FALSE)
curve(dnorm(x),col="red",add=TRUE)
```

4) Télécharger le fichier `dataforTP4` sur la page web `http://etienne.roquain.free.fr/teaching.html`. Charger des jeux de données simulées `Xobs1`et `Xobs2` à l'aide de la commande.

```
load("dataforTP4")
```

Est ce que ces données sont gaussiennes ? Pour chacun des jeux de données calculer la $p$-value du test par permutation à l'aide des questions précédentes.

5) Pour `Xobs2`, refaire la question précédente avec $B = 10^5$. Quel est l'avantage et l'inconvénient de la méthode de test par permutation ?

## Exercice 2 : Test multiple par permutation pour trouver les gènes différentiellement exprimés

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).

```r
library(sda)

data(singh2002)

prostate=singh2002
X=prostate$x
dim(X)
n=dim(X)[1]
m=dim(X)[2]
n0=sum(prostate$y==prostate$y[1]) # groupe sain
n1=n-n0 #groupe malade
```

Nous avons déjà vu que l'hypothèse gaussienne était peu vraisemblable. Aussi, *nous ne ferons pas l'hypothèse gaussienne ici.*

Aussi, une statistique de test raisonnable pour voir la différence entre les deux groupes pour `un` gène était donné par la fonction

```r
T=function(x) abs(t.test(x[1:n0],x[(n0+1):n],var.equal=TRUE)$stat)
```

1) Calculer le vecteur `stat` des valeurs des statistiques de tests pour les gènes. Faire le plot des statistiques.
2) On pourrait calculer les $p$-values de chaque gène par permutation et appliquer une correction de Bonferroni. Quel serait le principal inconvenient de cette méthode ? Quel $B$ faudrait-il choisir pour avoir un FWER contrôlé au niveau 5% ?
3) Calculer une fonction `maxT` qui a une matrice $M$ de taille $n \times m$ associe la valeur maximum des $T(M_j)$ où $M_1, \ldots, M_m$ désigne ses colonnes.
4) Que fait le programme suivant ?

```r
B=100
Sigma=matrix(sapply(1:B,function(b) sample(n)),n,B,byrow=FALSE)
maxTperm=sapply(1:B,function(b) maxT(X[Sigma[,b],]))
hist(maxTperm,freq=FALSE)
```

4) Calculer le nombre de rejets de la procédure de test multiple `maxT` par permutation au niveau $\alpha = 5\%$.