TP4 : Test multiple dans le modèle de bruit blanc gaussien

M2 Statistique, 2019-2020

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

Exercice 1 : Identification des 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).

```
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
```

1) Calculer et représenter le vecteur $Z = (Z_1, \dots, Z_m)$ de la façon suivante :

```
computestat=function(data) t.test(data[1:n0],data[(n0+1):n],var.equal=TRUE)$stat
Z=apply(X,2,computestat)
plot(Z,xlab="",ylab="",pch=19,col=gray(0.5))
```

En déduire une expression pour les p-values et les représenter.

2) Considérons $\alpha = 0.1$. Combien y-a t'il de p-values plus petites que α ? Commenter.

3) Calculer et représenter l'ensemble des p-values rejetées par la procédure de Bonferroni (au niveau $\alpha = 0.1$). Comment interpréter le résultat ?

4) Calculer et représenter l'ensemble des p-values rejetées par la procédure de Benjamini-Hochberg (au niveau $\alpha = 0.1$). Comment interpréter le résultat ?

Exercice 2 : Illustration de la procédure BH

Le but est de comprendre ce que fait l'interface disponible soit sur le web https://roquain.shinyapps.io/BH_illustration/, soit sur ma page web `BH_illustration.Rmd`.