Statistical modelling of complex data

Anne Gégout-Petit
Université de Lorraine, Faculté des Sciences, IECL, Inria Nancy

4 Lectures high-dimensional statistical analysis

1. Multiple testing issues
    - FWER: Family Wise Error Rate
    - False Discovery Rate
2. Model selection and assessment
    - Problematic, error
    - Criteria for linear model : AIC, BIC, Cp
    - Cross Validation and bootstrap method
    - Variable selection: subset
3. Regularization Methods for regression
    - Ridge Regression
    - Lasso method
    - Elastic-net
4. Reduction dimension methods
    - Principal Component Analysis
    - Partial Least Square regression
    - Sparse Methods
    - Discriminant Analysis version

Modelling complex data

1. Handling missing data
   - Framework, definition
   - Method using maximisation of the likelihood: EM algorithm
   - Imputation methods
   - Method for and with PCA
2. General Additive model
   - What is Additive Model?
   - What is GAM
   - Add smooth effect
   - Some tools for model selection
3. Network Inference
   - What is a network ...
   - Inference of a network
   - Graphical Gaussian Model

Organisation of a session

After 1h or more : we split into two groups to practice on your computer

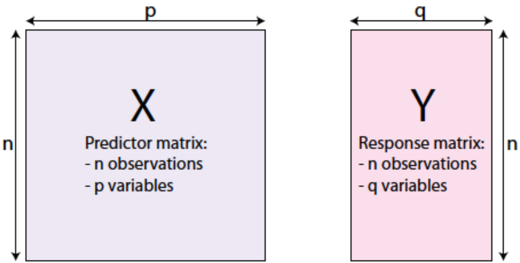
- You have to come with your laptop
- Some of the reports (code + pdf) must be submitted on Arche

Project, evaluation

1. Some reports of practical work are evaluated
2. Project at the end of the course : choose between
   - 2.1 Analysis of a complex dataset
     - 2.1.1 You find yourself your data and issue
     - 2.1.2 Validation of dataset and questions (by me !)
     - 2.1.3 One session (in November) for tutorial
   - 2.2 Study of a methodology
     - 2.2.1 Identify a research paper about a learning methodology
     - 2.2.2 Study the different steps
     - 2.2.3 Conduct a simulation study
     - 2.2.4 Apply on data
   - 2.3 For both of them
     - 2.3.1 You shall write a dissertation and the R programm
     - 2.3.2 Last session for the defences ...
     - 2.3.3 One note for the report
     - 2.3.4 One note for the defences
   - 2.4 December the 2nd: defense of the projects

High Dimensional Data: Challenge

Data Structure



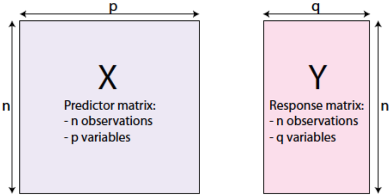
Identify which of the p covariates in X are associated with the outcome Y

Exemple : X are the omics data ($p = 20000$) and Y is the occurrence of a disease
Very frequently, we have $n < p$ and even $n \ll p$:

- More predictors than observations
- numerically intractable statistical inferences

High Dimensional Data: Challenge

Data Structure



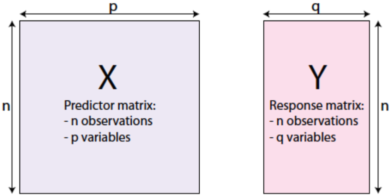
Identify which of the p covariates in X are associated with the outcome Y

If $n < p$, several possibilities

- tackle the link between Y and each of the covariates
  - How ?

High Dimensional Data: Challenge

Data Structure



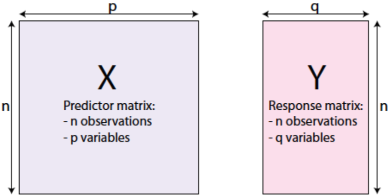
Identify which of the p covariates in X are associated with the outcome Y

If $n < p$, several possibilities

- tackle the link between Y and each of the covariates
  - How ?
  - by p statistical tests
  - by p regression with one regressor
  - by a multiple regression $\Rightarrow$ not possible without penalisation
- summarize X into a matrix of lower dimension
  - How ?

High Dimensional Data: Challenge

Data Structure



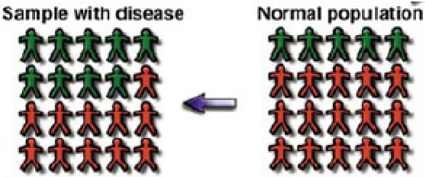
Identify which of the p covariates in X are associated with the outcome Y

If $n < p$, several possibilities

- tackle the link between Y and each of the covariates
  - How ?
  - by p statistical tests
  - by p regression with one regressor
  - by a multiple regression $\Rightarrow$ not possible without penalisation
- summarize X into a matrix of lower dimension
  - How ?
  - by Principal Component Analysis
  - by other factorial analysis
  - by "conducted" dimension reduction method
- Variable selection approach: find the best combination of covariates to predict Y

Example: Genome Wide Association Study (GWAS)

- Genetic variation associated to an univariate or multivariate disease phenotypes (Phenotypes: how geneticists spell Y)
  - univariate and binary Phenotype: Disease status
    - First successful GWAS in 2005: investigated patients age-related macular degeneration
    - a lot of variants has been found to various complex diseases: prostate or breast cancer, Crohn's disease, ...
    - more recent question about "personnalised medicine", what are the variants associated to the success of a treatment ?



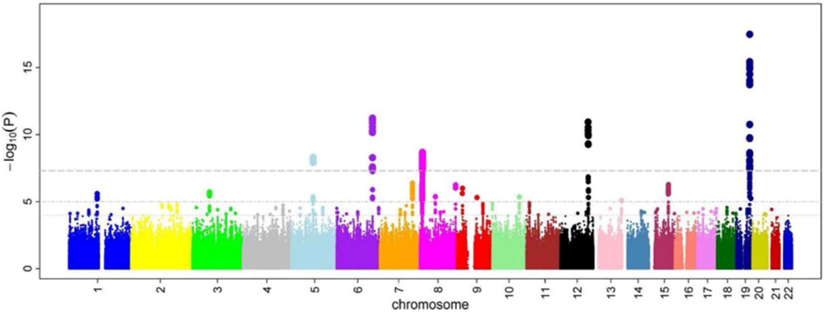
  - continuous phenotype univariate or multivariate: blood lipid (HDL, LDL cholesterol, triglyceride), blood pressure, ...
    - markers of cardiovascular disease
- Aim: identify the predictors associated to disease phenotypes

Tackle the link between Y and each of the covariates

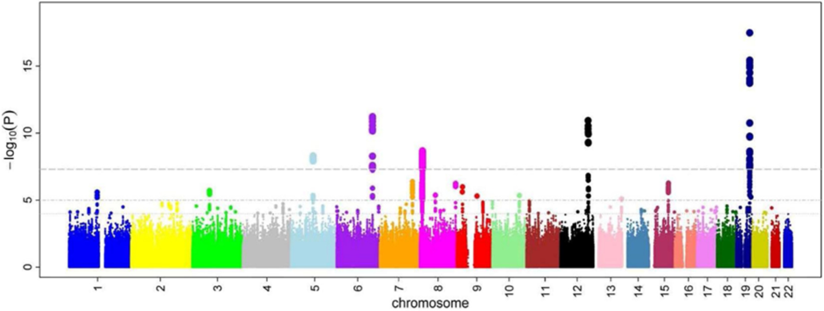
How ?

Tackle the link between Y and each of the covariates

Perform as many statistical tests as the number of covariates

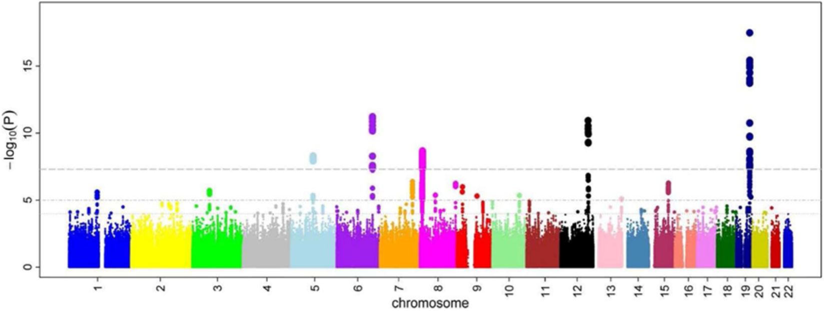


Tackle the link between Y and each of the covariates



Keep the covariates with p-value < threshold

Tackle the link between Y and each of the covariates



In typical omics dataset, the number of covariates $p > 30000$! $\Rightarrow$ multiple testing !

5 Lectures on Model Selection and high-dimensional statistical analysis

1. Today Multiple testing issues
   - FWER: Family Wise Error Rate
   - False Discovery Rate
2. Model selection and assessment
   - Problematic, error
   - Criteria for linear model : AIC, BIC, Cp
   - Cross Validation and bootstrap method
   - Variable selection: subset
3. Regularization Methods
   - Ridge Regression
   - Lasso method
   - Elastic-net
4. Reduction method
   - Principal Component Analysis
   - Partial Least Square regression
   - Sparse Methods
   - Discriminant Analysis version
5. General Additive model
   - What is Additive Model?
   - What is GAM
   - Add smooth effect
   - Some tools for model selection

Multiple testing

Aims

- Define the multiple testing issue and related concepts
- Methods for addressing multiple testing (FWER and FDR)
- Pratical implementation using R software

About a simple test

How it works ?

- You put an hypothesis that is called H_0

About a simple test

How it works ?

- You put an hypothesis that is called H_0
  typically H_0 : the mean of X_j is the same in population 1 and in population 2
- Find the distribution of a statistics T that measures something under H_0

About a simple test

How it works ?

- You put an hypothesis that is called H_0
  typically H_0 : the mean of X_j is the same in population 1 and in population 2
- Find the distribution of a statistics T that measures something under H_0
  typically ($T = \dfrac{\overline{X_1} - \overline{X_2}}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$ and under good hypotheses $T \simeq \mathcal{N}(0,1)$)

About a simple test

How it works ?

- You put an hypothesis that is called H_0
  typically H_0 : the mean of X_j is the same in population 1 and in population 2
- Find the distribution of a statistics T that measures something under H_0
  typically ($T = \dfrac{\overline{X_1} - \overline{X_2}}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$ and under good hypotheses $T \simeq \mathcal{N}(0,1)$)
- Choose a risk α , there are two equivalent ways to conclude, for this we have to compute the value t of T on a sample
  1. Reject if $t \in \Gamma_\alpha$ such that $P(T \in \Gamma_\alpha | H_0) = \alpha$
     typically ....

About a simple test

How it works ?

- You put an hypothesis that is called H_0
  typically H_0 : the mean of X_j is the same in population 1 and in population 2
- Find the distribution of a statistics T that measures something under H_0
  typically ($T = \dfrac{\overline{X_1} - \overline{X_2}}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$ and under good hypotheses $T \simeq \mathcal{N}(0,1)$)
- Choose a risk α , there are two equivalent ways to conclude, for this we have to compute the value t of T on a sample
  1. Reject if $t \in \Gamma_\alpha$ such that $P(T \in \Gamma_\alpha | H_0) = \alpha$
     typically ....
  2. Compute the p-value of the test : pv the smallest risk such that the value t is rejected and reject if $pv \leq \alpha$
     Typically ....

Errors associated to a test

	Decision	
”Truth” (unknown)	Reject H_0	Keep H_0
H_0 true	Incorrect decision	Correct decision
H_1 true	Correct decision	Incorrect decision

Errors associated to a test

	Decision	
”Truth” (unknown)	Reject H_0	Keep H_0
H_0 true	Incorrect decision Type I error	Correct decision
H_1 true	Correct decision	Incorrect decision

$$\alpha = P(\text{Type I error})$$

If we perform one test, we reject if the p-value$< \alpha$

Errors associated to a test

	Decision	
”Truth” (unknown)	Reject H_0	Keep H_0
H_0 true	Incorrect decision Type I error	Correct decision
H_1 true	Correct decision	Incorrect decision Type II error

$$\alpha = P(\text{Type I error}) \qquad \beta = P(\text{Type II error})$$

Errors associated to a test

"Truth" (unknown)	Decision Reject H_0	 Keep H_0
H_0 true	Incorrect decision	Correct decision
H_1 true	Correct decision Power	Incorrect decision

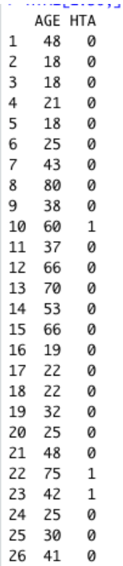
$\alpha = P(\text{Type I error})$ $\beta = P(\text{Type II error})$ $1 - \beta =$ "Power of the test"

Before to study multiple testing...

Some review about standard statistical test

Before to study multiple testing...

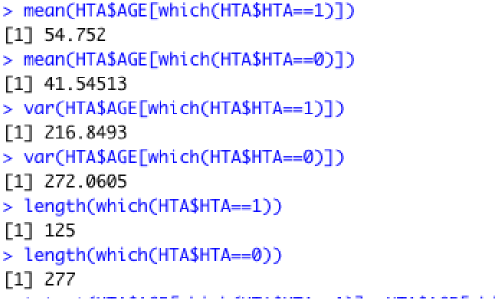
HTA = hypertension status



```
   AGE HTA
1   48   0
2   18   0
3   18   0
4   21   0
5   18   0
6   25   0
7   43   0
8   80   0
9   38   0
10  60   1
11  37   0
12  66   0
13  70   0
14  53   0
15  66   0
16  19   0
17  22   0
18  22   0
19  32   0
20  25   0
21  48   0
22  75   1
23  42   1
24  25   0
25  30   0
26  41   0
```

Before to study multiple testing...

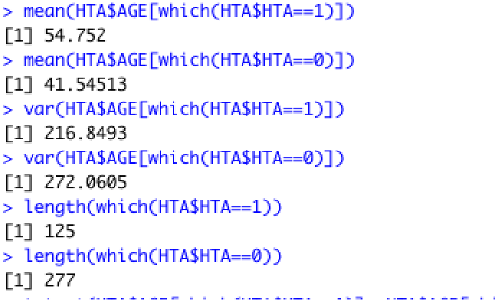
X quantitative, Y binary



```
> mean(HTA$AGE[which(HTA$HTA==1)])
[1] 54.752
> mean(HTA$AGE[which(HTA$HTA==0)])
[1] 41.54513
> var(HTA$AGE[which(HTA$HTA==1)])
[1] 216.8493
> var(HTA$AGE[which(HTA$HTA==0)])
[1] 272.0605
> length(which(HTA$HTA==1))
[1] 125
> length(which(HTA$HTA==0))
[1] 277
```

Before to study multiple testing...

X quantitative, Y binary

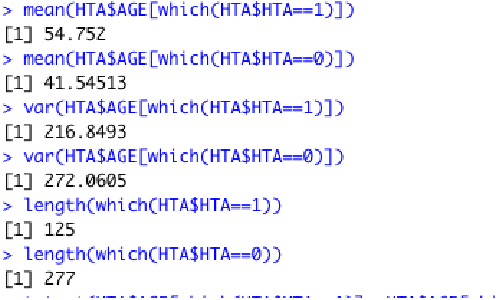


```
> mean(HTA$AGE[which(HTA$HTA==1)])
[1] 54.752
> mean(HTA$AGE[which(HTA$HTA==0)])
[1] 41.54513
> var(HTA$AGE[which(HTA$HTA==1)])
[1] 216.8493
> var(HTA$AGE[which(HTA$HTA==0)])
[1] 272.0605
> length(which(HTA$HTA==1))
[1] 125
> length(which(HTA$HTA==0))
[1] 277
```

Decision ? p-value ?

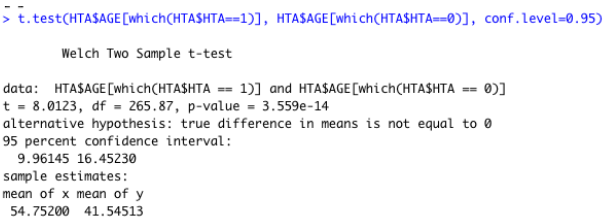
Before to study multiple testing...

X quantitative, Y binary



```
> mean(HTA$AGE[which(HTA$HTA==1)])
[1] 54.752
> mean(HTA$AGE[which(HTA$HTA==0)])
[1] 41.54513
> var(HTA$AGE[which(HTA$HTA==1)])
[1] 216.8493
> var(HTA$AGE[which(HTA$HTA==0)])
[1] 272.0605
> length(which(HTA$HTA==1))
[1] 125
> length(which(HTA$HTA==0))
[1] 277
```

Student-Test comparison of the means of two independent samples

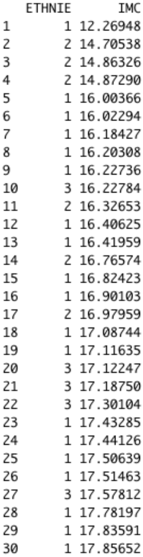


```
> t.test(HTA$AGE[which(HTA$HTA==1)], HTA$AGE[which(HTA$HTA==0)], conf.level=0.95)

        Welch Two Sample t-test

data:  HTA$AGE[which(HTA$HTA == 1)] and HTA$AGE[which(HTA$HTA == 0)]
t = 8.0123, df = 265.87, p-value = 3.559e-14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  9.96145 16.45230
sample estimates:
mean of x mean of y
 54.75200  41.54513
```

Before to study multiple testing...

Some review about standard statistical test

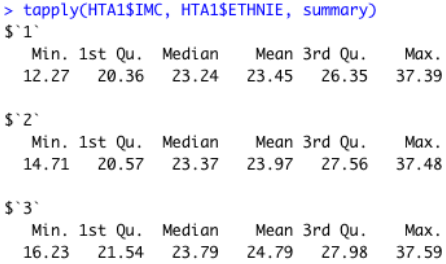
Before to study multiple testing...



```
   ETHNIE      IMC
1       1 12.26948
2       2 14.70538
3       2 14.86326
4       2 14.87290
5       1 16.00366
6       1 16.02294
7       1 16.18427
8       1 16.20308
9       1 16.22736
10      3 16.22784
11      2 16.32653
12      1 16.40625
13      1 16.41959
14      2 16.76574
15      1 16.82423
16      1 16.90103
17      2 16.97959
18      1 17.08744
19      1 17.11635
20      3 17.12247
21      3 17.18750
22      3 17.30104
23      1 17.43285
24      1 17.44126
25      1 17.50639
26      1 17.51463
27      3 17.57812
28      1 17.78197
29      1 17.83591
30      1 17.85652
```

Before to study multiple testing...

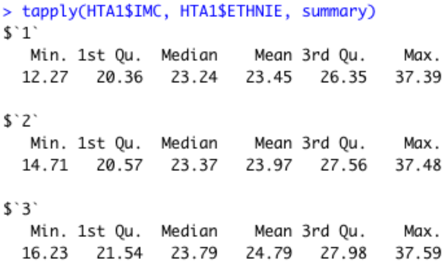
X quantitative, Y quali with more than 2 levels



```
> tapply(HTA1$IMC, HTA1$ETHNIE, summary)
$`1`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  12.27   20.36   23.24   23.45   26.35   37.39 

$`2`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  14.71   20.57   23.37   23.97   27.56   37.48 

$`3`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  16.23   21.54   23.79   24.79   27.98   37.59 
```

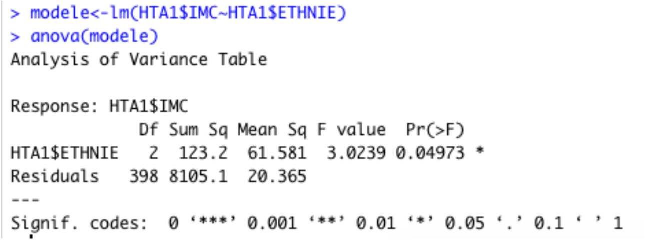
Before to study multiple testing...

X quantitative, Y quali with more than 2 levels



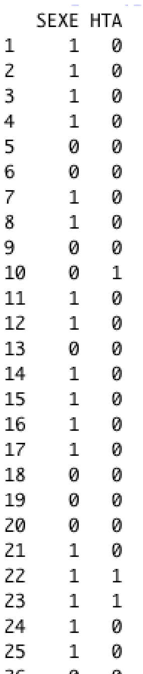
```
> tapply(HTA1$IMC, HTA1$ETHNIE, summary)
$`1`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  12.27   20.36   23.24   23.45   26.35   37.39 

$`2`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  14.71   20.57   23.37   23.97   27.56   37.48 

$`3`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  16.23   21.54   23.79   24.79   27.98   37.59 
```

Anova



```
> modele<-lm(HTA1$IMC~HTA1$ETHNIE)
> anova(modele)
Analysis of Variance Table

Response: HTA1$IMC
             Df Sum Sq Mean Sq F value  Pr(>F)  
HTA1$ETHNIE   2  123.2  61.581  3.0239 0.04973 *
Residuals   398 8105.1  20.365                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

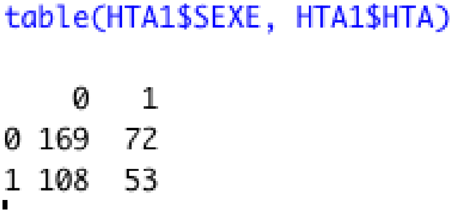
Before to study multiple testing...



```
    SEXE HTA
1      1   0
2      1   0
3      1   0
4      1   0
5      0   0
6      0   0
7      1   0
8      1   0
9      0   0
10     0   1
11     1   0
12     1   0
13     0   0
14     1   0
15     1   0
16     1   0
17     1   0
18     0   0
19     0   0
20     0   0
21     1   0
22     1   1
23     1   1
24     1   0
25     1   0
```

Before to study multiple testing...

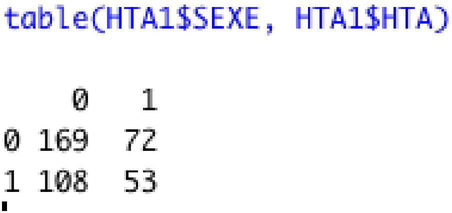
X qualitative, Y binary



```
table(HTA1$SEXE, HTA1$HTA)

     0   1
0 169  72
1 108  53
```

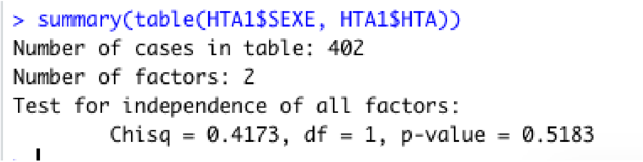
Before to study multiple testing...

X qualitative, Y binary



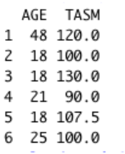
```
table(HTA1$SEXE, HTA1$HTA)

     0   1
0 169  72
1 108  53
```

Test ? Decision ? p-value ? **Chi square test**



```
> summary(table(HTA1$SEXE, HTA1$HTA))
Number of cases in table: 402
Number of factors: 2
Test for independence of all factors:
        Chisq = 0.4173, df = 1, p-value = 0.5183
```

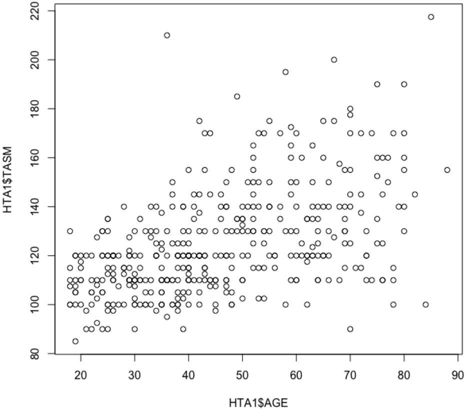
Before to study multiple testing...



```
  AGE  TASM
1  48 120.0
2  18 100.0
3  18 130.0
4  21  90.0
5  18 107.5
6  25 100.0
```

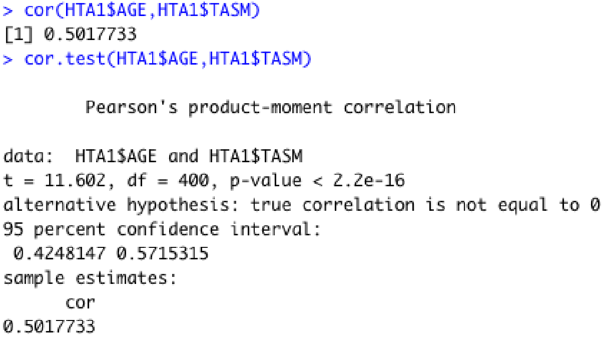
Before to study multiple testing...

X and Y qualitative



Before to study multiple testing...

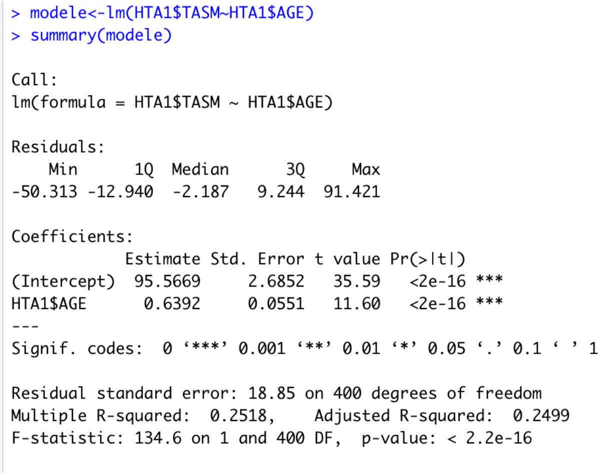
Test for the correlation



```
> cor(HTA1$AGE,HTA1$TASM)
[1] 0.5017733
> cor.test(HTA1$AGE,HTA1$TASM)

        Pearson's product-moment correlation

data:  HTA1$AGE and HTA1$TASM
t = 11.602, df = 400, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4248147 0.5715315
sample estimates:
      cor
0.5017733
```

Before to study multiple testing...

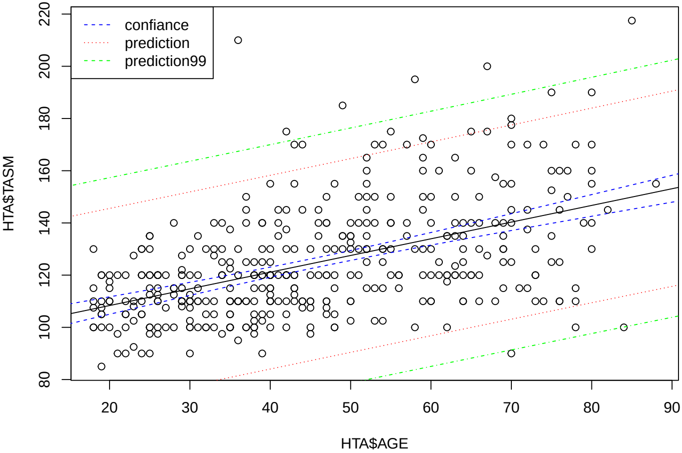
Linear regression



```
> modele<-lm(HTA1$TASM~HTA1$AGE)
> summary(modele)

Call:
lm(formula = HTA1$TASM ~ HTA1$AGE)

Residuals:
    Min      1Q  Median      3Q     Max 
-50.313 -12.940  -2.187   9.244  91.421 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  95.5669     2.6852   35.59   <2e-16 ***
HTA1$AGE      0.6392     0.0551   11.60   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.85 on 400 degrees of freedom
Multiple R-squared:  0.2518,	Adjusted R-squared:  0.2499 
F-statistic: 134.6 on 1 and 400 DF,  p-value: < 2.2e-16
```

Before to study multiple testing...

Linear regression



Multiple testing: what is the problem ?

We test the link between each X_j and Y by testing H_0^j, $(1 \leq j \leq p)$

	Decision	
"Truth" (unknown)	Reject $\mathrm{H}_0^{\mathrm{j}}$	Keep $\mathrm{H}_0^{\mathrm{j}}$
$\mathrm{H}_0^{\mathrm{j}}$ true	Incorrect decision X_j false positive	Correct decision X_j true negative
$\mathrm{H}_1^{\mathrm{j}}$ true	Correct decision X_j true positive	Incorrect decision X_j false negative

Multiple testing: what is the problem ?

We test the link between each X_j and Y by testing H_0^j, $(1 \leq j \leq p)$

"Truth" (unknown)	Decision Reject $\mathrm{H}_0^{\mathrm{j}}$	 Keep $\mathrm{H}_0^{\mathrm{j}}$
$\mathrm{H}_0^{\mathrm{j}}$ true	Incorrect decision X_j false positive	Correct decision X_j true negative
$\mathrm{H}_1^{\mathrm{j}}$ true	Correct decision X_j true positive	Incorrect decision X_j false negative

If the $p = 20\ 000$ test are independent

- If for each test $\alpha = 0.05$, expected number of false positive test: $p \times \alpha = 1000$
- If we expect only one false positive, we have to choose $\alpha = 0.05/p = 2.5 \times 10^{-6}$
- How to control the number of false positive ?

Multiple testing: what is the problem ?

If we perform p independent tests, what is the probability to have at least one false positive ?

$$P(\text{ type I error for one test}) = \alpha$$

Multiple testing: what is the problem ?

If we perform p independent tests, what is the probability to have at least one false positive ?

$$\begin{aligned} P(\text{ type I error for one test}) &= \alpha \\ P(\text{no type I error for one test}) &= 1-\alpha \end{aligned}$$

Multiple testing: what is the problem ?

If we perform p independent tests, what is the probability to have at least one false positive ?

$$
\begin{aligned}
P(\text{ type I error for one test}) &= \alpha \\
P(\text{no type I error for one test}) &= 1-\alpha \\
P(\text{no type I error in } p \text{ tests}) &= (1-\alpha)^p \\
P(\text{at least 1 type I error in } p \text{ tests}) &= 1-(1-\alpha)^p
\end{aligned}
$$

Multiple testing: what is the problem ?

If we perform p independent tests, what is the probability to have at least one false positive ?

$$\begin{aligned}
P(\text{ type I error for one test}) &= \alpha \\
P(\text{no type I error for one test}) &= 1-\alpha \\
P(\text{no type I error in } p \text{ tests}) &= (1-\alpha)^p \\
P(\text{at least 1 type I error in } p \text{ tests}) &= 1-(1-\alpha)^p
\end{aligned}$$

Compute it for $p = 2, 5, 10, 20$...

Probablility to have one error

If we perform *p* independent tests, what is the probability to have at least one false positive ?

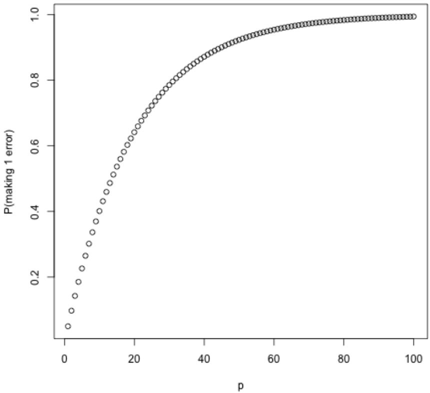


Figure: Probability to have at least one type I error according to the number *p* of independent statistical tests ($\alpha = 0.05$ for each test)

Counting false and true decisions

Suppose we perform p tests : $H_0^1, H_0^2, \dots H_0^p$, We denote $p_0 = \#$ true hypothesis and $R = \#$ of rejected hypothesis

We have to control some numbers !

	Significant	Not significant	Total
Null true	V	U	p_0
Alternative true	S	T	$p - p_0$
Total	R	p-R	p

Counting false and true decisions

Suppose we perform p tests : $H_0^1, H_0^2, \dots H_0^p$, We denote $p_0 = \#$ true hypothesis and $R = \#$ of rejected hypothesis

	Significant	Not significant	Total
Null true	V	U	p_0
Alternative true	S	T	$p - p_0$
Total	R	p-R	p

V is the number of type I errors or false positive Which quantities are known if we

perform p test from real data ?

Multiple testing Aim

- "Adjusting p-values for the number of hypotheses tests performed" means controlling the Type I error
- Very active area of statistics
- Many different methods with the same goal but with fundamentally different ways

Multiple testing Aim

- "Adjusting p-values for the number of hypotheses tests performed" means controlling the Type I error
- How ? according to $\alpha = P(\text{at least 1 type I error in } p \text{ tests})$, we modify the level α_j of the test H_0^j or in a equivalent manner the p-value pv_j.

Main FWER control procedures

FWER, Family Wise Error Rate, we want to control
$\alpha = P(\text{at least 1 type I error in } p \text{ tests})$,

- Single step approaches (Bonferroni, Sidak, ...)
- Bonferroni, for a overall error α, level for H_0^j is $\alpha_j = \frac{\alpha}{p}$
  - In an equivalent manner, we compare $p \times pv_j$ to α, $\tilde{pv}_j^{Bonf} = min(p \times pv_j, 1)$ is called the Bonferroni adjusted pvalue and is to be compared to α

$$pv_j < \alpha/p \Leftrightarrow \tilde{pv}_j^{Bonf} < \alpha$$

.

Main FWER control procedures

FWER, Family Wise Error Rate, we want to control
$\alpha = P(\text{at least 1 type I error in } p \text{ tests})$,

- Single step approaches (Bonferroni, Sidak, ...)
- Bonferroni, for a overall error α, level for H_0^j is $\alpha_j = \frac{\alpha}{p}$
  - In an equivalent manner, we compare $p \times pv_j$ to α, $\tilde{pv}_j^{Bonf} = min(p \times pv_j, 1)$ is called the Bonferroni adjusted pvalue and is to be compared to α

$$pv_j < \alpha/p \Leftrightarrow \tilde{pv}_j^{Bonf} < \alpha$$

.

- Sidak: for a overall error α we reject H_0^i at level $\alpha_j = 1 - (1 - \alpha)^{1/p}$
  - Exercice: compute the corresponding Sidak adjusted p-value.

Main FWER control procedures

FWER, Family Wise Error Rate, we want to control
$\alpha = P$(at least 1 type I error in p tests),

- Single step approaches (Bonferroni, Sidak, ...)
- Bonferroni, for a overall error α, level for H_0^j is $\alpha_j = \frac{\alpha}{p}$
  - In an equivalent manner, we compare $p \times pv_j$ to α, $\tilde{pv}_j^{Bonf} = min(p \times pv_j, 1)$ is called the Bonferroni adjusted pvalue and is to be compared to α

$$pv_j < \alpha / p \Leftrightarrow \tilde{pv}_j^{Bonf} < \alpha$$

.

- Sidak: for a overall error α we reject H_0^i at level $\alpha_j = 1 - (1 - \alpha)^{1/p}$
  - Exercice: compute the corresponding Sidak adjusted p-value.
- These two procedures are very conservative ! High probability of type II error of not rejecting the general null hypothesis when important effects exist
- Very contre-intuitive: a results for one covariates depends on the number of tests!
- The adjusting procedure does not depend on the p-value

FWER sequential adjustment

- Sequential method means that the adjustement depends on the order of the p-value
- Simplest sequential method is Holms Method
  1. Order the unadjusted p-values such that $pv_{(1)} \leq pv_{(2)} \leq \ldots \leq pv_{(p)}$ and also the associated hypothesis $H_0^{(1)}\ H_0^{(2)}\ \ldots\ H_0^{(p)}$
  2. For a given significance level α, let k be the minimal index such that $pv_{(k)} > \frac{\alpha}{p+1-k}$
  3. Reject $H_0^{(1)}\ H_0^{(2)}\ \ldots\ H_0^{(k-1)}$ and do not reject $H_0^{(k)}\ H_0^{(2)}\ \ldots\ H_0^{(p)}$
  4. The corresponding adjusted p-values are

FWER sequential adjustment

- Sequential method means that the adjustement depends on the order of the p-value
- Simplest sequential method is Holms Method
  1. Order the unadjusted p-values such that $pv_{(1)} \leq pv_{(2)} \leq \ldots \leq pv_{(p)}$ and also the associated hypothesis $H_0^{(1)}\ H_0^{(2)}\ \ldots\ H_0^{(p)}$
  2. For a given significance level α, let k be the minimal index such that $pv_{(k)} > \frac{\alpha}{p+1-k}$
  3. Reject $H_0^{(1)}\ H_0^{(2)}\ \ldots\ H_0^{(k-1)}$ and do not reject $H_0^{(k)}\ H_0^{(2)}\ \ldots\ H_0^{(p)}$
  4. The corresponding adjusted p-values are

$$\tilde{pv}_{(j)}^{Holms} = min\left((p-j+1) \times pv_{(j)},\ 1\right)$$

- The point here is that we don't multiply every pv_j by the same factor p : it is sequential or stepwise

Not making ANY type I Errors ?

- FWER is appropriate when you want to guard against ANY false positives
- However, in many cases (particularly in genomics) we can tolerate a moderate number of False Positive
- In these cases, the more relevant quantity to control is the false discovery rate (FDR)

What is False Discovery Rate ?

	Significant	Not significant	Total
Null true	V	U	p_0
Alternative true	S	T	$p - p_0$
Total	R	p-R	p

What is random in this table ?
What is unknown in this table ?

What is False Discovery Rate ?

	Significant	Not significant	Total
Null true	V	U	p_0
Alternative true	S	T	$p - p_0$
Total	R	p-R	p

- FDR $= \mathbb{E}[V/R]$. It is the expected proportion of False Positive among the significant tests (R)
- The adjustment aim is to ensure that FDR is upper bounded by a desired value
- Main FDR procedure: Benjamini Hochberg, it is stepwise
- FDR *vs*. FWER control: FDR is less stringent than FWER
  - FWER controls $P(V \geq 1)$.
  - FDR control: over p experiments the average of $FP/R \leq \alpha$
  - FDR control may be preferred in an exploratory context

First task : Estimating p_0

	Significant	Not significant	Total
Null true	V	U	p_0
Alternative true	S	T	$p - p_0$
Total	R	p-R	p

What is observed in this table ?

First task : Estimating p_0

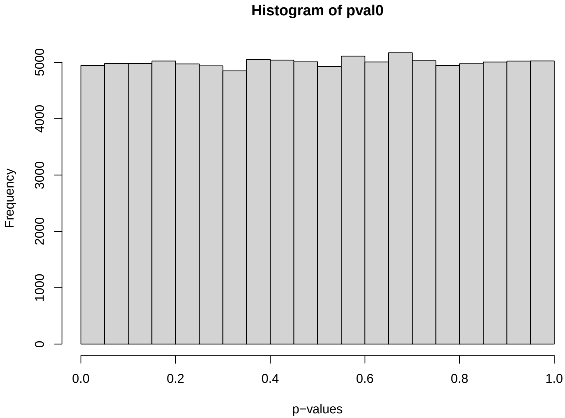
	Significant	Not significant	Total
Null true	V	U	p_0
Alternative true	S	T	$p - p_0$
Total	R	p-R	p

p tests $\Rightarrow$ p p-values !

Use properties of p-values under H_0 to estimate p_0

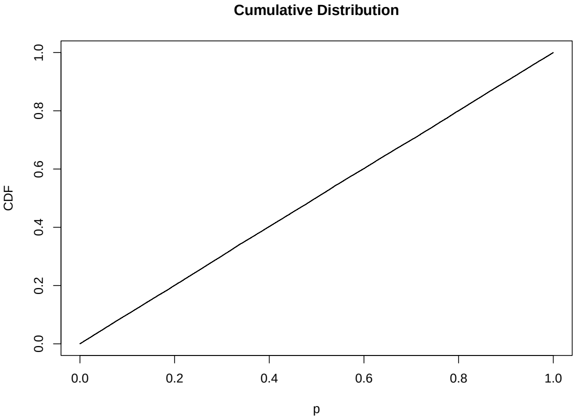
Estimating p_0

Under H_0 p-values are uniformly distributed on $[0, 1]$.



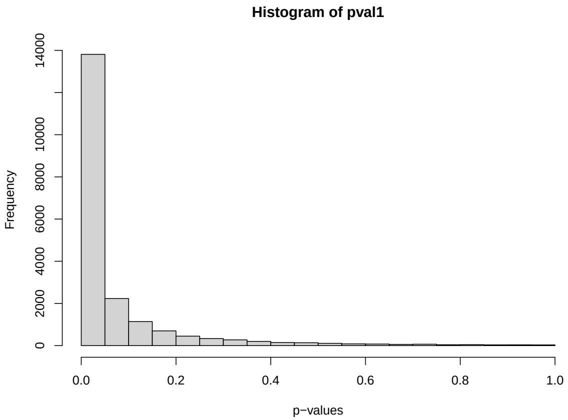
Estimating p_0

Under H_0 p-values are uniformly distributed on $[0, 1]$.



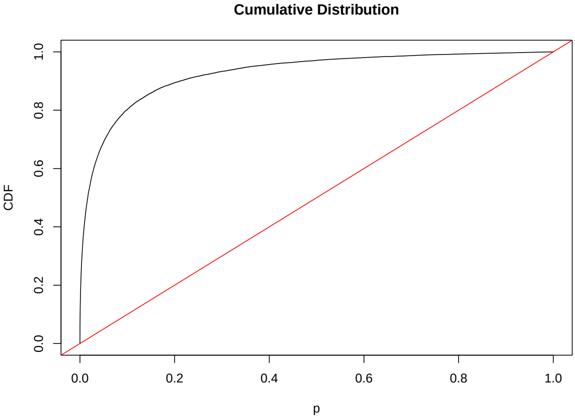
Estimating p_0

Under the alternative hypothesis H_1, p-values are skewed towards 0



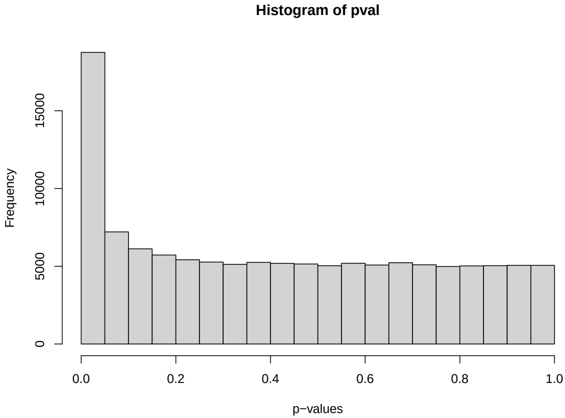
Estimating p_0

Under the alternative hypothesis H_1, p-values are skewed towards 0



Estimating p_0

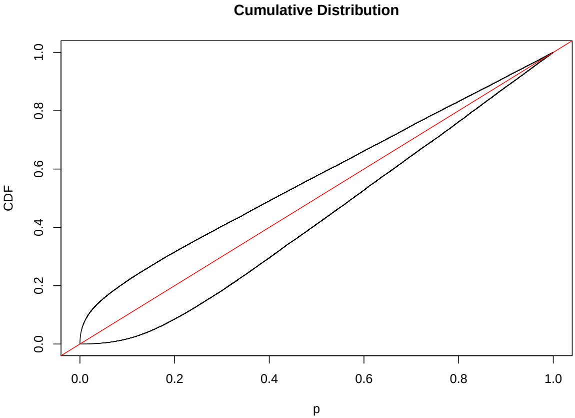
If we have a mixture of the two hypotheses p_0 under H_0 and $p - p_0$ under H_1:



How to disentangle the null from the alternatives ?

Estimating p_0

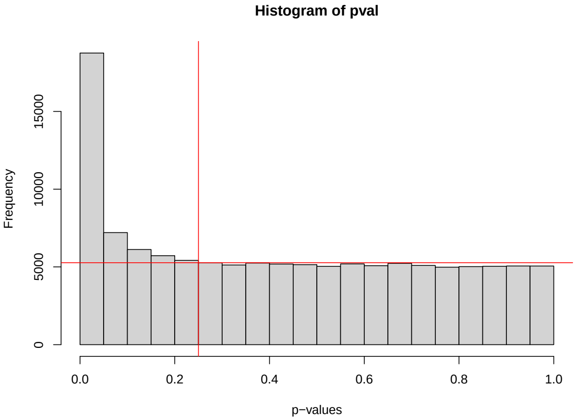
If we have a mixture of the two hypotheses p_0 under H_0 and $p - p_0$ under H_1:



How to disentangle the null from the alternatives ?

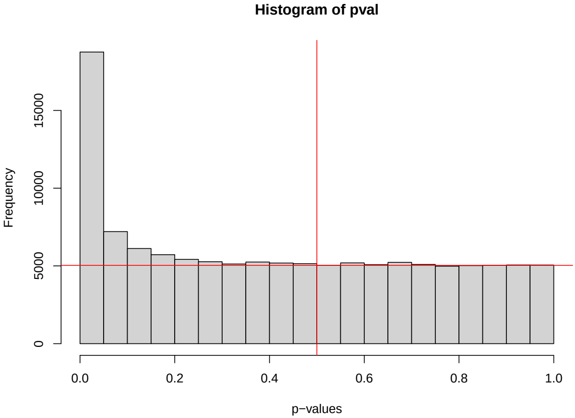
Estimating p_0

We locate the threshold λ of uniformity: here for p-values greater than $\lambda = 0.25$, we assume they mostly represent observations from null hypothesis



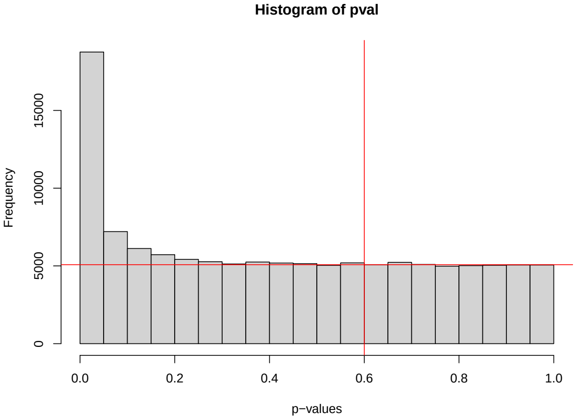
Estimating p_0

We locate the threshold λ of uniformity: here for p-values greater than $\lambda = 0.5$, we assume they mostly represent observations from null hypothesis



Estimating p_0

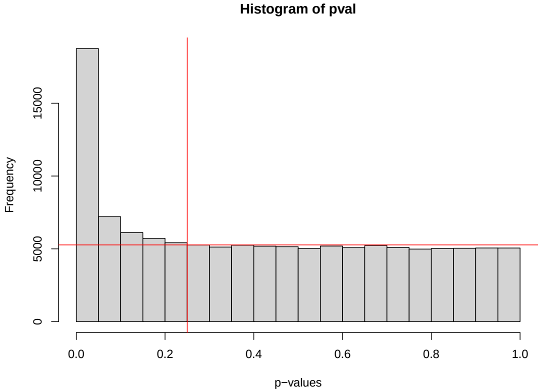
We locate the threshold λ of uniformity: here for p-values greater than $\lambda = 0.5$, we assume they mostly represent observations from null hypothesis



Estimating p_0

One can estimate p_0 by

One can estimate p_0 by $$\hat{p}_0 = \frac{\#\{p_i > \lambda\}}{(1-\lambda)}$$



Simulations : $p_0 = 100000, p = 120000$
Estimations : for $\lambda = 0.25$, $\hat{p}_0 = 102805$, for $\lambda = 0.5$, $\hat{p}_0 = 101760$, for $\lambda = 0.6$, $\hat{p}_0 = 101843$,

After estimating p_0, choose to reject the $\hat{p}_0$ smallest p-values

In our exemples

	Significant	Not significant	Total
Null true	4284	95716	100000
Alternative true	13956	6044	20000
Total	18240	101760	120000

Here FDR=4284/18240=0.235

Second task : Estimating p_0 but control FDR

Benjamini & Hochberg (1995) (BH) proposed a step-wise method for controlling FDR

1. Compare the largest p-value among the p with the chosen specified significance level α: if
$$pv_{(p)} > \alpha,$$
then do not reject the corresponding hypothesis $H_0^{(p)}$
2. Compare the second one to a modified threshold:
$$pv_{(p-1)} > \alpha \times (p-1)/p, \quad \Rightarrow \text{do not reject} \quad H_0^{(p-1)}$$
3. $$pv_{(p-2)} > \alpha \times (p-2)/p, \quad \Rightarrow \text{do not reject} \quad H_0^{(p-2)}$$
4. ...
5. Stop when p-value is lower than the modified threshold, all other null hypotheses (with smaller p-values) are rejected

BH adjusted p-values

- FDR is being controlled
- If the hypotheses are independent, the set of decisions verifies

$$FDR = \mathbb{E}\left[\frac{V}{R}\right] \leq (p_0/p)\alpha \leq \alpha$$

Property of BH method

$$\underbrace{H_0^{(1)} \ldots H_0^{(j^*)}}_{\text{rejected}} \underbrace{H_0^{(j^*+1)} \ldots H_0^{(p)}}_{\text{not rejected}}$$

$$j^* = min\{j \text{ such that } pv_{(j+1)} > \alpha \frac{j+1}{p}\}$$

$$j^* = min\{j \text{ such that } \frac{p \cdot pv_{(j+1)}}{j+1} > \alpha\}$$

$$\underbrace{pv_{(1)} \leq \ldots \leq pv_{(j^*)}}_{< \frac{\alpha j^*}{p}} \leq \underbrace{pv_{(j^*+1)} \leq \ldots \leq pv_{(p)}}_{\text{each} pv_{(k)} > \frac{\alpha k}{p} \geq \frac{\alpha j*}{p}}$$

Adjusted p-values could be given by $p\frac{p_{(k)}}{j^*}$, but $j*$ depends on α and on the p-values ! Software compute the adjusted Benjamini-Hochberg p-values by

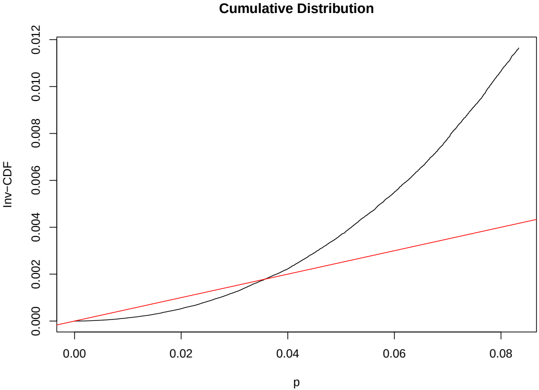
$$\tilde{pv}_{(j)}^{BH} = \min\left(\min_{i \geq j}(p * pv_{(i)}/i),\ 1\right)$$

Property of BH method

$$j^* = min\{j \text{ such that } pv_{(j+1)} > \alpha\frac{j+1}{p}\}$$

$$\underbrace{pv_{(1)} \leq \ldots \leq pv_{(j^*)}}_{< \frac{\alpha j^*}{p}} \leq \underbrace{pv_{(j^*+1)} \leq \ldots \leq pv_{(p)}}_{\text{each} pv_{(k)} > \frac{\alpha k}{p} \geq \frac{\alpha j*}{p}}$$

Threshold is given by the intersection between the inverse of the CDF and line of



gradient α

Results for $\alpha = 0.1$

In our exemples

	Significant	Not significant	Total
Null true	549	79541	100000
Alternative true	8328	11672	20000
Total	9024	110976	120000

Here FDR=549/9024=0.060

Results for $\alpha = 0.2$

In our exemples

	Significant	Not significant	Total
Null true	1813	78187	100000
Alternative true	8087	11672	20000
Total	14169	105831	120000

Here FDR=1813/14169=0.128

A Bayesian approach to FDR

Storey (2000)

$$pFDR = \mathbb{E}[\left(\frac{V}{R}|R > 0\right)]$$

- Assume i.i.d. statistics $T_1 \dots T_p$ and rejection region Γ.
- Define Z_j equals 0 if H_0^j is true and 1 otherwise

  $$T_j|Z_j \simeq (1 - Z_j)F_0 + Z_jF_1$$

  for some distribution F_0 and F_1.
- Letting $P(Z_j = 0) = \pi_0$, we have

  $$T_j \simeq \pi_0 F_0 + (1 - \pi_0)F_1$$

Storey showed

$$pFDR(\Gamma) = P(Z_j = 0|T_j \in \Gamma)$$

posterior probability that the null hypothesis is true given than test statistics falls in the rejection region for the test.

positive False Discovery Rate (pFDR)

How to estimate the pFDR ?

$$\begin{aligned} pFDR &= \mathbb{P}(H_0 | T \in \Gamma) \\ &= \frac{\mathbb{P}(T \in \Gamma | H_0)\mathbb{P}(H_0)}{\mathbb{P}(T \in \Gamma)} \end{aligned}$$

- $\mathbb{P}(T \in \Gamma | H_0)$ (proba of type 1 error risk when coosing Γ)
- $\mathbb{P}(T \in \Gamma) = R/p$ (proportion of hypotheses rejected)
- $\mathbb{P}(H_0) = \pi_0$ (to be estimated from the data by $\hat{\pi}_0 = \frac{\#\{p_i > \lambda\}}{p(1-\lambda)}$ for instance)

The qvalues

qvalue, Definition: The q-value of a test T_j is defined to be the smallest pFDR over all rejection regions that reject Tj .

$$\begin{aligned} pFDR &= \frac{\mathbb{P}(T_j \in \Gamma | H_0)\mathbb{P}(H_0)}{\mathbb{P}(T_j \in \Gamma)} \\ q_j &= \hat{\pi}_0 pv_j p/R \end{aligned}$$

- Note similarity between the adjusted p-values using the BH method
- q-values are not linear w.r.t. p-values because of R (number of rejected nulls, it is a $R(\Gamma)$, it changes with Γ or with the risk

References

Review of frequentist methods, controlling error rates:

Dudoit, Shaffer & Boldrick (2003), Statistical Science 18, p 71.
http://www.stat.berkeley.edu/sandrine/
Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS Genet 2:e190.

About FDR

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Series B, 57:289-300.
Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. Ann Stat 29:1165-1188.

Estimating pFDR:

Storey JD. 2002. A direct approach to false discovery rates. J R Stat Soc Series B Stat Methodol 64:479-498.

Storey JD. 2003. The positive false discovery rate: A Bayesian interpre- tation and the q-value. Ann Stat 31:2013-2035