# Tutoriel sur les tests multiples (et au-delà)

Timothée Flutre

# Tutoriel sur les tests multiples (et au-delà)

Timothée Flutre

11/03/2016

# Philo de modélisation statistique en trois phrases

- Box (1987): “Essentially, all models are wrong, but some are useful.”

# Philo de modélisation statistique en trois phrases

- Box (1987): “Essentially, all models are wrong, but some are useful.”

- Kass (2011): “When we use a statistical model to make a statistical inference, we implicitly assert that the variation exhibited by data is captured reasonably well by the statistical model, so that the theoretical world corresponds reasonably well to the real world.”

# Philo de modélisation statistique en trois phrases

- Box (1987): "Essentially, all models are wrong, but some are useful."
- Kass (2011): "When we use a statistical model to make a statistical inference, we implicitly assert that the variation exhibited by data is captured reasonably well by the statistical model, so that the theoretical world corresponds reasonably well to the real world."
- Berger & Berry (1988): "It is not possible to provide an absolutely objective answer [to a statistical test]; the strength of the evidence will depend on the person interpreting the data."

# Notations

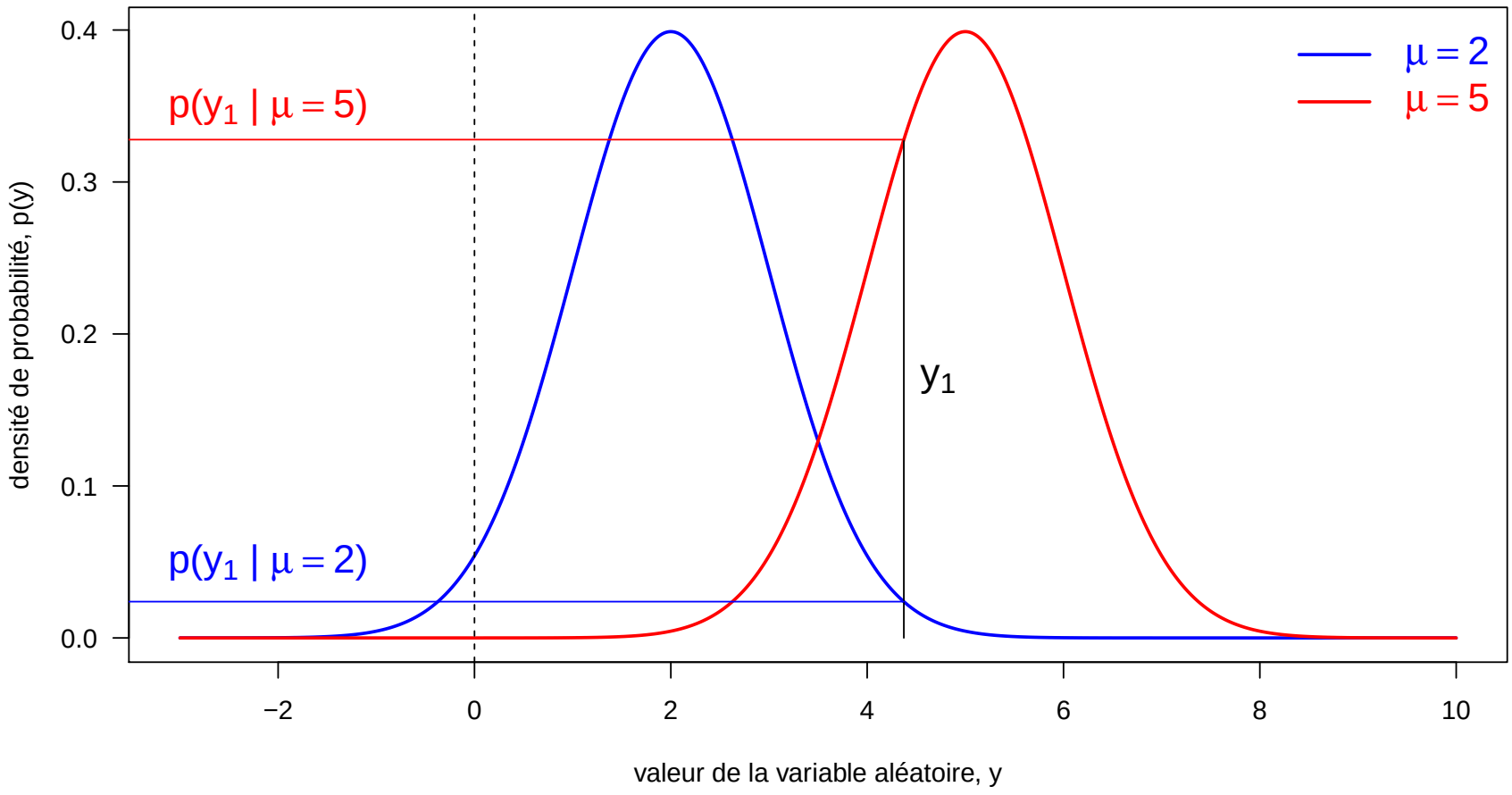
- données: variables observées, lettres romaines, $y, x$
- paramètres: variables non-observées, lettres grecques, $\theta, \beta$
- ensembles: majuscules, $\mathcal{D}$ (données) et $\Theta$ (paramètres)
- vecteurs: en gras, $\boldsymbol{y}$, $\boldsymbol{\beta}$

- vraisemblance: proba des données sachant les paramètres, mais fonction des paramètres, $\mathcal{L}(\Theta) = p(\mathcal{D} \mid \Theta)$
- maximum de vraisemblance: pour estimer les paramètres, $\hat{\theta} = \text{argmax}_\theta \, \mathcal{L} \quad \Leftrightarrow \quad \frac{\partial \mathcal{L}}{\partial \theta}(\hat{\theta}) = 0$

# Concrètement...

Simuler $Y \sim \mathcal{N}(\mu, \sigma^2)$ avec $\sigma^2 = 1$ renvoie $y_1 = 4.37$: que vaut $\mu$ ?

Maximiser $\mathcal{L}(\mu) = p(y_1|\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y_1-\mu)^2}{2}\right)$



# Exemple 1

On veut tester si le SNP chr3_716254_A_C est associé avec le caractère "rendement" au sein de $N$ variétés.

- données $\mathcal{D}$: $(x_i, y_i)$ génotype et phénotype de la variété $i$
- paramètres $\Theta$: $\mu$ moyenne globale, $\beta$ effet du génotype, $\sigma^2$ variance des erreurs
- hypothèses: ...
- vraisemblance $\mathcal{L}(\Theta) = p(\mathcal{D}|\Theta)$:

$$\forall i,\ y_i = \mu + \beta x_i + \epsilon_i \text{ avec } \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- estimer par maximum de vraisemblance l'effet du génotype au SNP sur le phénotype, $\hat{\beta}$, et son erreur standard, $s$

# Exemple 2

On veut tester si l'expression du gène MAP3 dans la variété Inadur change selon qu'on irrigue ou pas.

- données $\mathcal{D}$: $(x_i, y_i)$ indicateur d'irriguation et expression pour le plant $i$
- paramètres $\Theta$: $\mu$ moyenne globale, $\beta$ effet de l'irriguation, $\sigma^2$ variance des erreurs
- hypothèses: ...
- vraisemblance $\mathcal{L}(\Theta) = p(\mathcal{D}|\Theta)$:

$$\forall i,\ y_i = \mu + \beta x_i + \epsilon_i \text{ avec } \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- estimer par maximum de vraisemblance l'effet de l'irriguation sur l'expression du gène, $\hat{\beta}$, et son erreur standard, $s$

# Estimation de paramètre

Procédure générique: maximum de vraisemblance

- vraisemblance: $\mathcal{L}(\Theta) = \prod_{i=1}^{N} p(y_i|x_i, \mu, \beta, \sigma)$

# Estimation de paramètre

Procédure générique: maximum de vraisemblance

- vraisemblance: $\mathcal{L}(\Theta) = \prod_{i=1}^{N} p(y_i|x_i, \mu, \beta, \sigma)$
- maximisation: $\frac{\partial \mathcal{L}}{\partial \beta}(\hat{\beta}) = 0$

# Estimation de paramètre

Procédure générique: maximum de vraisemblance

- vraisemblance: $\mathcal{L}(\Theta) = \prod_{i=1}^{N} p(y_i|x_i, \mu, \beta, \sigma)$
- maximisation: $\frac{\partial \mathcal{L}}{\partial \beta}(\hat{\beta}) = 0$
- estimation: $\hat{\beta} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$

# Estimation de paramètre

Procédure générique: maximum de vraisemblance

- vraisemblance: $\mathcal{L}(\Theta) = \prod_{i=1}^{N} p(y_i|x_i, \mu, \beta, \sigma)$
- maximisation: $\frac{\partial \mathcal{L}}{\partial \beta}(\hat{\beta}) = 0$
- estimation: $\hat{\beta} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$

- estimateur fréquentiste: $B = \frac{\sum_i (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_i (x_i - \bar{x})^2}$

# Estimation de paramètre

Procédure générique: maximum de vraisemblance

- vraisemblance: $\mathcal{L}(\Theta) = \prod_{i=1}^{N} p(y_i|x_i, \mu, \beta, \sigma)$
- maximisation: $\frac{\partial \mathcal{L}}{\partial \beta}(\hat{\beta}) = 0$
- estimation: $\hat{\beta} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$

- estimateur fréquentiste: $B = \frac{\sum_i (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_i (x_i - \bar{x})^2}$
- $E[B] = \beta$ et $V[B] = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} \Rightarrow s^2 = \frac{1}{N-2} \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (x_i - \bar{x})^2}$

# Estimation de paramètre

Procédure générique: maximum de vraisemblance

- vraisemblance: $\mathcal{L}(\Theta) = \prod_{i=1}^{N} p(y_i|x_i, \mu, \beta, \sigma)$
- maximisation: $\frac{\partial \mathcal{L}}{\partial \beta}(\hat{\beta}) = 0$
- estimation: $\hat{\beta} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$

- estimateur fréquentiste: $B = \frac{\sum_i (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_i (x_i - \bar{x})^2}$
- $E[B] = \beta$ et $V[B] = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} \Rightarrow s^2 = \frac{1}{N-2} \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (x_i - \bar{x})^2}$

- hypothèse: erreurs Normales $\Rightarrow$ $Y_i \sim \mathcal{N}$ $\Rightarrow$ $B \sim \mathcal{N}$

# Test d'hypothèse

- hypothèse nulle, $H_0$: "$\beta = 0$"

# Test d'hypothèse

- hypothèse nulle, $H_0$: "$\beta = 0$"
- statistique de test (Wald): $Z|H_0 = \frac{B}{\sqrt{Var(B)}} \sim \mathcal{N}(0,1)$

# Test d'hypothèse

- hypothèse nulle, $H_0$: "$\beta = 0$"
- statistique de test (Wald): $Z|H_0 = \frac{B}{\sqrt{Var(B)}} \sim \mathcal{N}(0,1)$
- réalisation: $z = \frac{\hat{\beta}}{s}$

# Test d'hypothèse

- hypothèse nulle, $H_0$: "$\beta = 0$"
- statistique de test (Wald): $Z|H_0 = \frac{B}{\sqrt{Var(B)}} \sim \mathcal{N}(0,1)$
- réalisation: $z = \frac{\hat{\beta}}{s}$
- probabilité critique (*p value*): $p = P(Z \geq z|H_0)$

# Test d'hypothèse

- hypothèse nulle, $H_0$: "$\beta = 0$"
- statistique de test (Wald): $Z|H_0 = \frac{B}{\sqrt{Var(B)}} \sim \mathcal{N}(0,1)$
- réalisation: $z = \frac{\hat{\beta}}{s}$
- probabilité critique (*p value*): $p = P(Z \geq z|H_0)$
- Wald: approxime LRT mais équivalent asymptotiquement

# Test d'hypothèse

- hypothèse nulle, $H_0$: "$\beta = 0$"
- statistique de test (Wald): $Z|H_0 = \frac{B}{\sqrt{Var(B)}} \sim \mathcal{N}(0,1)$
- réalisation: $z = \frac{\hat{\beta}}{s}$
- probabilité critique (*p value*): $p = P(Z \geq z|H_0)$
- Wald: approxime LRT mais équivalent asymptotiquement
- Z-score: perte d'info de $(\hat{\beta}, s)$ à $z$

# Simuler un petit jeu de données

```r
set.seed(1859)
N <- 100
x <- rbinom(n=N, size=2, prob=0.3)
mu <- 4
pve <- 0.4 # = var(x beta) / var(y)
sigma <- 1
(beta <- sigma * sqrt(pve / ((1 - pve) * var(x))))
```

```
## [1] 1.338481
```

```r
e <- rnorm(n=N, mean=0, sd=sigma)
y <- mu + beta * x + e
```

# Faire l'inférence (built-in)

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6384 -0.6207 -0.1253  0.6976  2.4843
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.1229     0.1331  30.965  < 2e-16 ***
## x             1.0450     0.1639   6.376 5.98e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 0.9948 on 98 degrees of freedom
## Multiple R-squared:  0.2932, Adjusted R-squared:  0.286
## F-statistic: 40.66 on 1 and 98 DF,  p-value: 5.975e-09
```

# Faire l'inférence (custom, estimation)

```
(beta.hat <- ((t(x) %*% y - N * mean(x) * mean(y)) /
                  (t(x) %*% x - N * mean(x)^2))[1,1])
```

```
## [1] 1.045039
```

```
(sigma.hat <- sqrt((1/(N-2) * sum(res$residuals^2))))
```

```
## [1] 0.9947526
```

```
(se.beta.hat <- sqrt(sigma.hat^2 /
                         (t(x) %*% x - N * mean(x)^2)[1,1]))
```

```
## [1] 0.1638911
```

# Faire l'inférence (custom, test)

```
(z.score <- beta.hat / se.beta.hat)
```

```
## [1] 6.376424
```

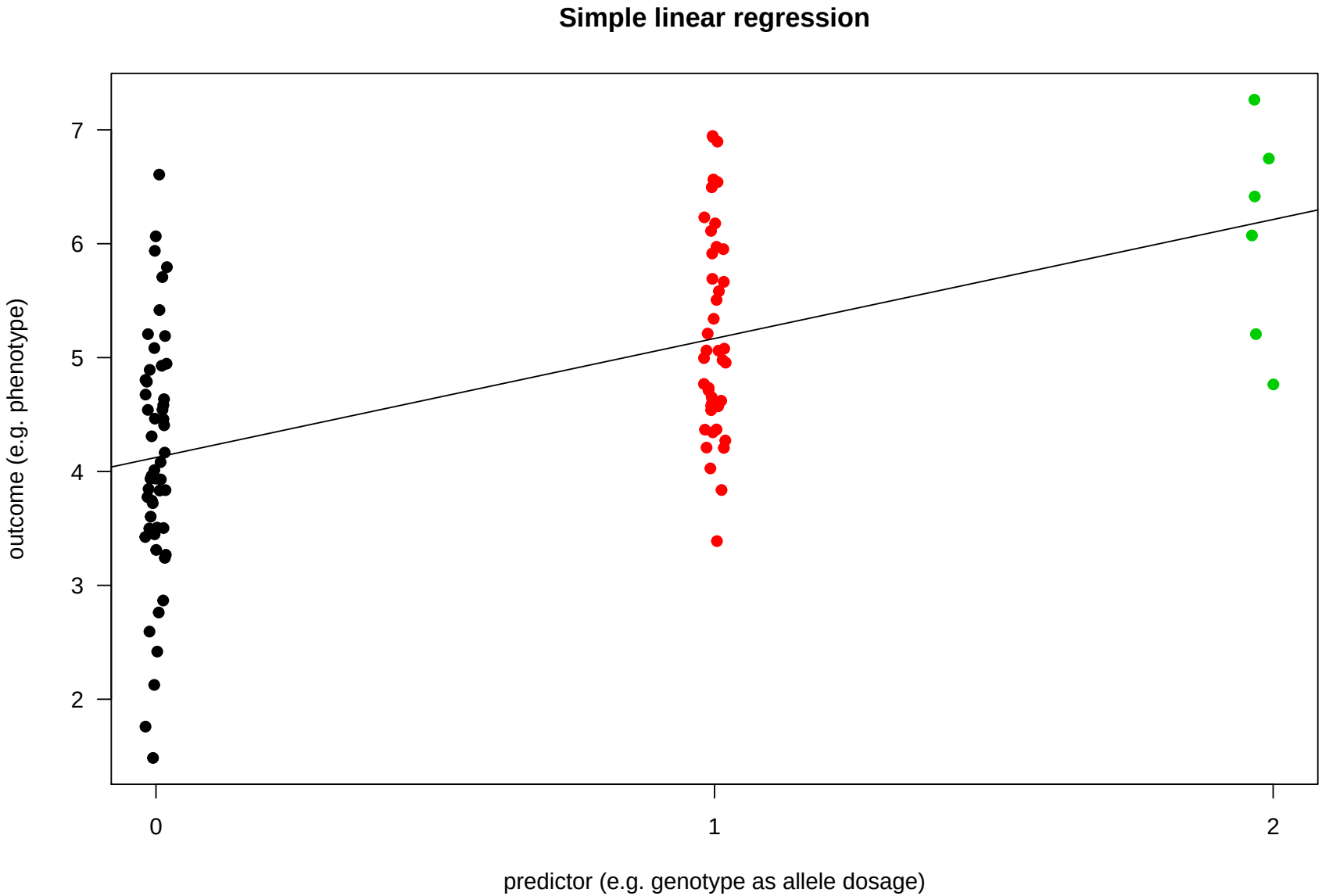
```
(pvalue <- 2 * pt(q=z.score, df=N-2, lower.tail=FALSE))
```

```
## [1] 5.975296e-09
```

```
(pvalue <- 2 * pnorm(q=z.score, mean=0, sd=1, lower.tail=FA
```

```
## [1] 1.812709e-10
```

# Visuellement



# Significativité de l'association

Si l'hypothèse nulle est fausse, on s'attend à ce que la proba critique soit faible, donc on rejette $H_0$ si $p \leq$ seuil, mais lequel?

Plusieurs cas possibles:

| | garder $H_0$ | rejeter $H_0$ |
|---|---|---|
| $H_0$ vraie | VN | FP |
| $H_0$ fausse | FN | VP |

On veut généralement limiter la proba, notée $\alpha$, d'avoir un faux positif (erreur de type I).

Comme, sous $H_0$, la proba critique suit une loi Uniforme sur $[0, 1]$, on a donc: $P(p \leq \text{seuil} | H_0) = \text{seuil}$.

$\Rightarrow$ on choisit de rejeter $H_0$ si $p \leq \alpha$, par ex le fameux 5%

# Point de vue bayésien

facteur de bayes: $\mathrm{BF} = \frac{P(\mathcal{D}|H_0)}{P(\mathcal{D}|H_1)} = \frac{\int p(\Theta_0)\, p(\mathcal{D}|\Theta_0)\, \mathrm{d}\Theta_0}{\int p(\Theta_1)\, p(\mathcal{D}|\Theta_1)\, \mathrm{d}\Theta_1}$

- garder $H_0$ si $\frac{P(H_0|\mathcal{D})}{P(H_1|\mathcal{D})} = \mathrm{BF}\, \frac{P(H_0)}{P(H_1)} < \frac{\mathrm{coût}_{II}}{\mathrm{coût}_{I}}$
- difficultés: intégration, prior, seuil

Idée (Johnson, Wakefield, Wen & Stephens): remplacer la vraisemblance $\boldsymbol{y}|\Theta$ par $\hat{\beta}|\beta \sim \mathcal{N}(\beta, s^2)$

- $BF \approx ABF = \sqrt{\frac{s_0^2+s^2}{s^2}} \exp\left(-\frac{z^2}{2}\frac{s_0^2}{s_0^2+s^2}\right)$ avec $\beta \sim \mathcal{N}(0, s_0^2)$

Pour un $N$ donné et un $s_0^2$ peu informatif, choisir $P(H_0)$ (ex. 0.5) et $\frac{\mathrm{coût}_{II}}{\mathrm{coût}_{I}}$ (ex. 1) permet de choisir le seuil sur le Z-score, et donc sur la proba critique, *seuil qui dépend maintenant de la puissance du test (via nb d'échantillons, N), quel que soit le nombre de tests...*

# Une analyse typique de génomique

1. génotyper $N$ individus à $P$ marqueurs, et phénotyper ces individus; ou bien mesurer l'expression de $P$ gènes chez $N$ individus avec ou sans traitement

# Une analyse typique de génomique

1. génotyper $N$ individus à $P$ marqueurs, et phénotyper ces individus; ou bien mesurer l'expression de $P$ gènes chez $N$ individus avec ou sans traitement

2. pour chaque $j \in \{1, \ldots, P\}$, inférer par maximum de vraisemblance: estimation de l'effet $\hat{\beta}_j$ et son erreur standard $s_j$

# Une analyse typique de génomique

1. génotyper $N$ individus à $P$ marqueurs, et phénotyper ces individus; ou bien mesurer l'expression de $P$ gènes chez $N$ individus avec ou sans traitement

2. pour chaque $j \in \{1, \ldots, P\}$, inférer par maximum de vraisemblance: estimation de l'effet $\hat{\beta}_j$ et son erreur standard $s_j$

3. les transformer en scores standardisés: $z_j = \frac{\hat{\beta}_j}{s_j}$

# Une analyse typique de génomique

1. génotyper $N$ individus à $P$ marqueurs, et phénotyper ces individus; ou bien mesurer l'expression de $P$ gènes chez $N$ individus avec ou sans traitement

2. pour chaque $j \in \{1, \ldots, P\}$, inférer par maximum de vraisemblance: estimation de l'effet $\hat{\beta}_j$ et son erreur standard $s_j$

3. les transformer en scores standardisés: $z_j = \frac{\hat{\beta}_j}{s_j}$

4. calculer les probabilités critiques, $p_j$, via $Z_j | H_0 \sim \mathcal{N}(0, 1)$

# Cas possibles (tableau)

| | garder $H_0$ | rejeter $H_0$ | |
|---|---|---|---|
| $H_0$ vraie | $VN$ | $FP$ | $P_0$ |
| $H_0$ fausse | $FN$ | $VP$ | $P_1$ |
| | | $R$ | $P$ |

où maintenant $FP$ est une variable contenant le nombre de tests correspondant à des faux-positifs

# Problème

Avec la même procédure que précédemment, le nombre de faux positifs augmente linéairement avec le nombre de tests...

Par exemple, même si $H_0$ est toujours fausse ($P_0 = P$) et $\alpha = 5\%$:

- $P = 500 \Rightarrow E[FP] = 25$
- $P = 1000 \Rightarrow E[FP] = 50$
- $P = 2000 \Rightarrow E[FP] = 100$

# Family-Wise Error rate (FWER)

A contrôler, par exemple via la procédure de Bonferroni:

$$
\begin{aligned}
FWER|\mathcal{H}_0 = \Pr(FP \geq 1|\mathcal{H}_0) &= \Pr(\bigcup_{j=1}^{P}\{p_j \leq \alpha_j|H_{0j}\}) \\
&\leq \sum_j \Pr(p_j \leq \alpha_j|H_{0j}) \\
&\leq \sum_j \alpha_j \ \leq \ \alpha \ \ \text{si } \forall j \ \alpha_j \leq \alpha
\end{aligned}
$$

- en pratique: `R> p.adjust(pvalues, "bonferroni")`
- FWER: critère (très) stringent (surtout si $P$ large)
- Bonferroni: d'autant plus conservatif que tests corrélés

# False Discovery rate (FDR)

$FDR = E[FP/R]$ et, par définition, $FDR = 0$ si $R = 0$

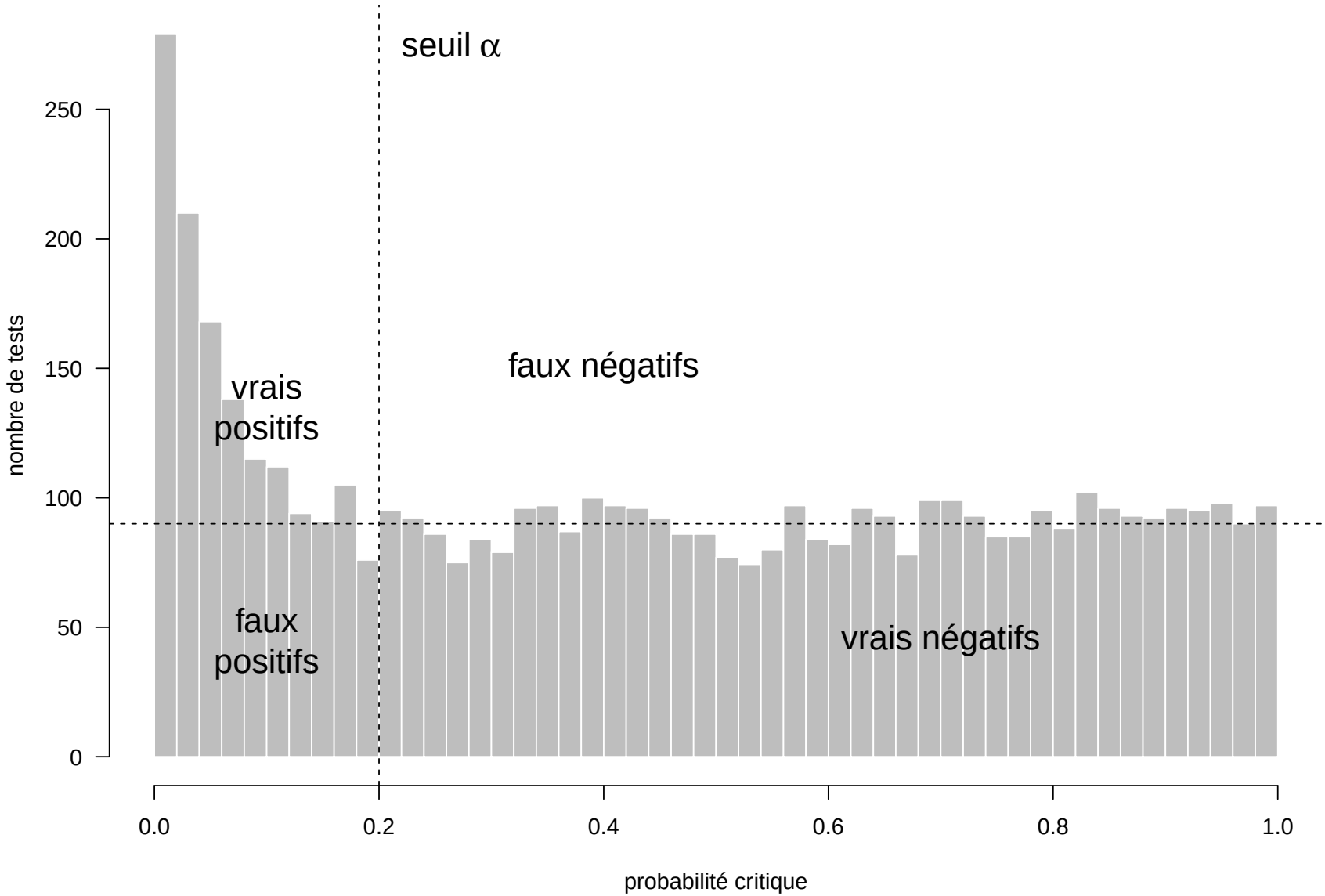
A contrôler, par exemple via la procédure de Benjamini-Hochberg:

- en pratique: `R> p.adjust(pvalues, "bh")`

Remarquez: $FDR = \Pr(R > 0)\ E[FP/R|R > 0]$

- problème: contrôler le FDR peut se faire en diminuant $\Pr(R > 0)$ et non $E[FP/R|R > 0]$ ...
- solution: contrôler le positive FDR: $pFDR = E[FP/R|R > 0]$

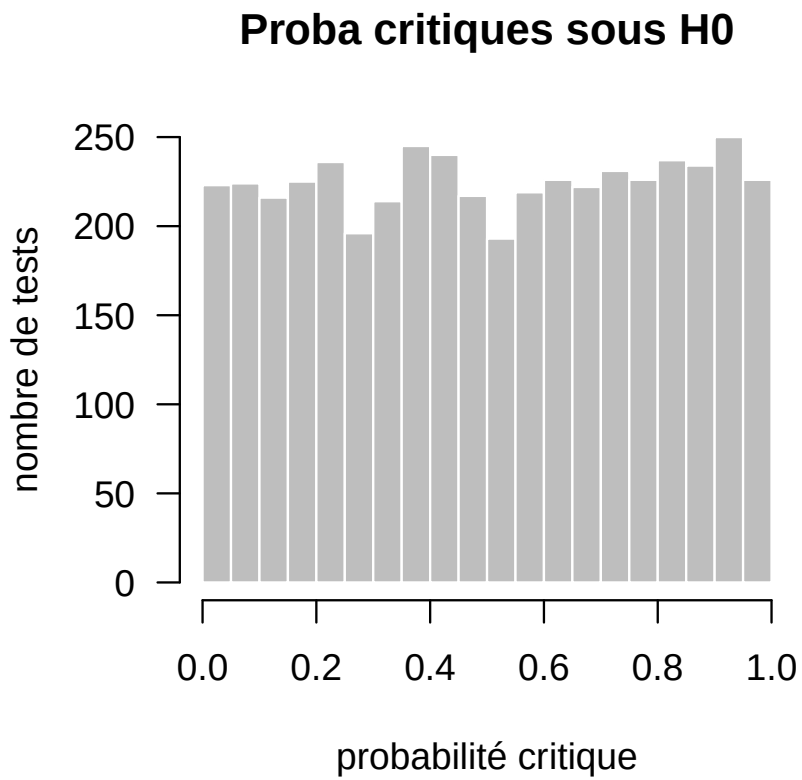
# Cas possibles (graphique)

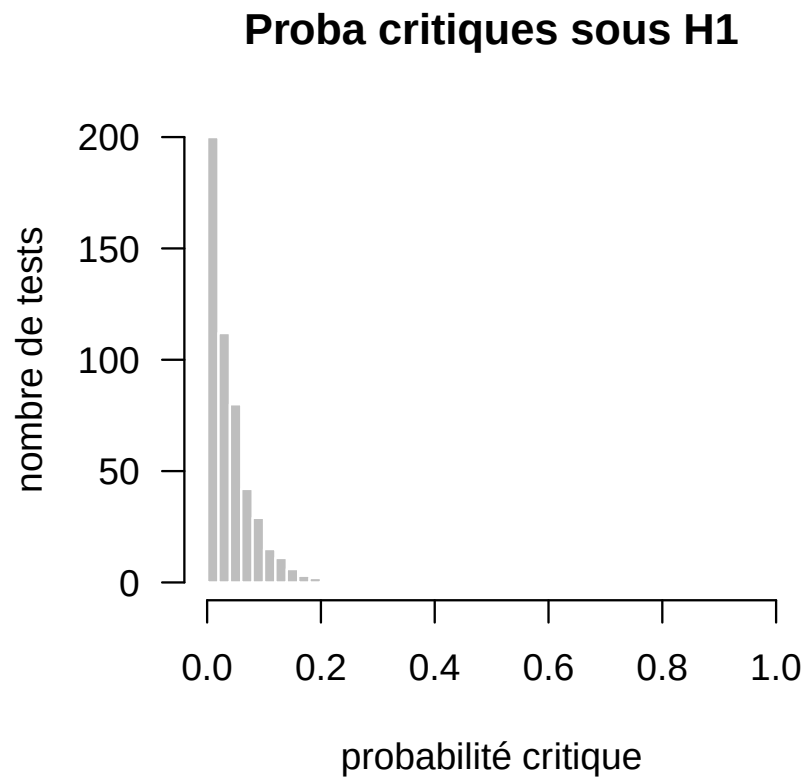


# pFDR et modèle de mélange

$$\forall j,\ p_j = \pi_0\, \mathcal{U}_{[0,1]} + (1-\pi_0)\, f_1$$

Exemple avec $\pi_0 = 0.9$:



# Procédure de Storey

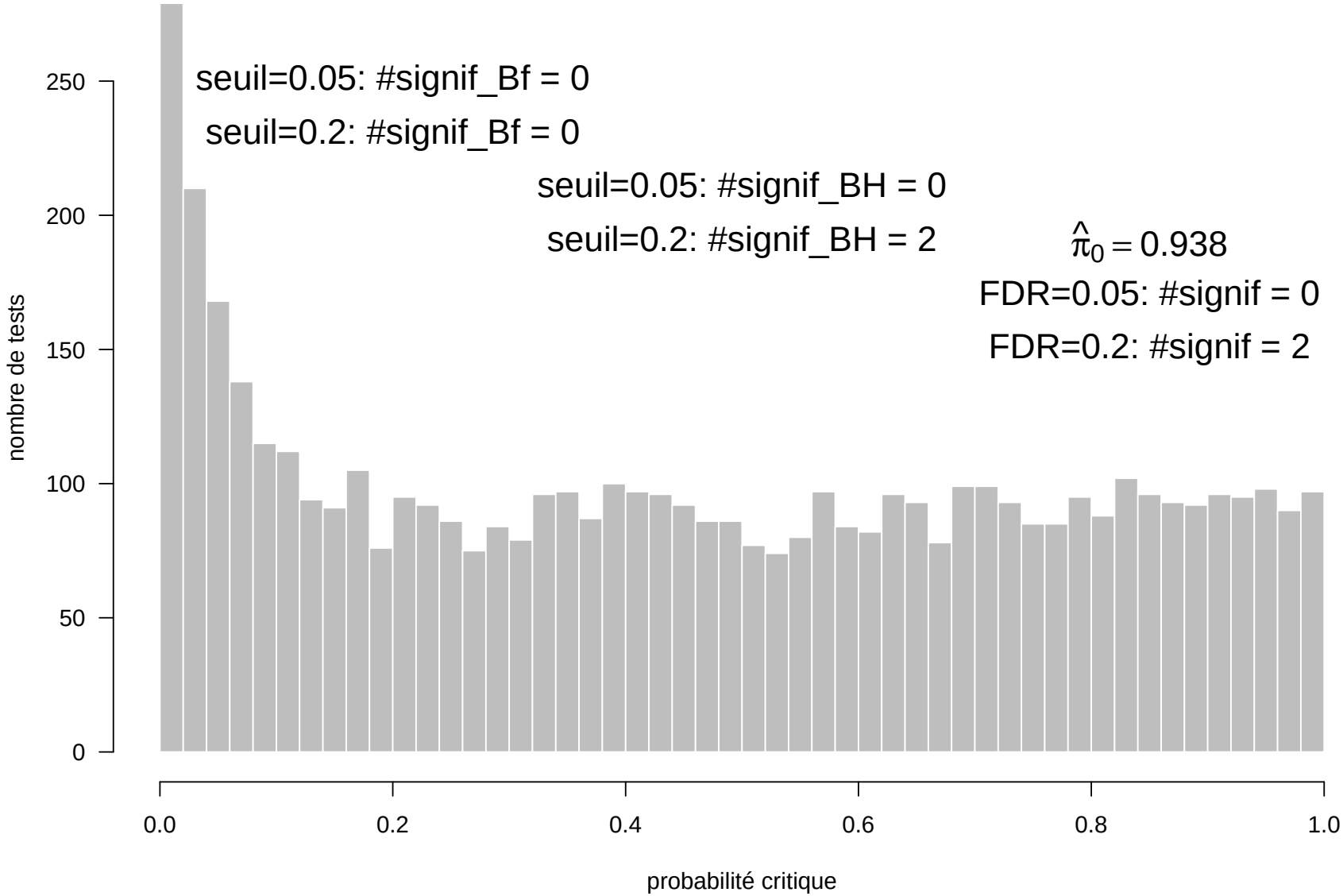
$$\widehat{pFDR}(\alpha) = \hat{\pi}_0 \frac{\alpha P}{\#\{p_j \leq \alpha\}}$$

- $\hat{\pi}_0$: calculer avec les proba critiques proches de 1

Si on ne veut pas fixer le seuil $\alpha$ par avance, on peut calculer une $q$-value par test, cad le pFDR pour tous les tests aussi *ou plus* significatifs que le test en question (fréquentiste):

- en pratique (paquet à installer): `R> qvalue(pvalues)`

# Application sur l'exemple



# local FDR (lfdr)

Proba qu'une certaine découverte est fausse à un seuil donné:

$lfdr_j = \Pr(\beta_j = 0 | z_j)$

- via le modèle de mélange en bayésien: $lfdr_j = \frac{\pi_0 f(z_j|\beta_j)}{f(z_j)}$

En pratique:

- $p_j$ transformée: locfdr (Efron, mais archive CRAN)
- $z_j$: mixfdr (Muralidharan, aussi archive CRAN)
- $\hat{\beta}_j, s_j$: ashr (Stephens, dépôt GitHub)

# False Sign rate (FSR)

Gelman & Tuerlinckx (2000): "we do not believe that $\beta = 0$ is a reasonable possibility for continuous parameters"

- type S error: wrongly identifying the sign of $\beta$ with confidence

Pr(type S error | claim with confidence)

$= \Pr(\text{sign}(\beta_j) \neq \text{sign}(\hat{\beta}_j) \mid 0 \notin \hat{\beta}_j \pm 1.96\, s_j)$

# Erreurs corrélées

Vraisemblance: $\forall j,\ \boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{x}_j\beta_j + Z\boldsymbol{u} + \boldsymbol{\epsilon}$

- $\boldsymbol{u} \sim \mathcal{N}(\mathbf{0}, \sigma_a^2 K)$
- $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$
- $Cov[\boldsymbol{u}, \boldsymbol{\epsilon}] = 0$

Utiliser le logiciel GEMMA (Zhou & Stephens, 2012):

- pour chaque SNP: $\hat{\beta}_j$ et $s_j$, statistique de Wald, p-value
- q-value, etc

Sinon le paquet R QTLRel.

# Analyser tous les SNPs dans le même modèle

Vraisemblance: $\boldsymbol{y} = \mathbf{1}\mu + X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ avec $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$

Prior: $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma^2_\beta I)$

Exemple de la taille chez l'homme:

- Yang et coll. (Nat Genet, 2010)
- Wood et coll. (Nat Genet, 2014)

Essayer le logiciel GCTA (maintenu par le labo de Peter Visscher).

Sinon le paquet R rrBLUP.

# Sélectionner les SNPs

Vraisemblance: $\boldsymbol{y} = \mathbf{1}\mu + X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ avec $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$

Prior: $\boldsymbol{\beta} \sim \pi_0\, \delta_0 + (1 - \pi_0)\, \mathcal{N}(\mathbf{0}, \sigma_\beta^2 I)$

$\Rightarrow \mathit{lfdr}_j = \Pr(\beta_j = 0 \,|\, \boldsymbol{y}, X)$

- BVSR: Guan & Stephens (Ann Appl Stat, 2011); comparaison avec LASSO; logiciel piMASS
- étendu par BSLMM: Zhou, Carbonetto & Stephens (PLoS Genet, 2013); logiciel GEMMA

# Tester différentes combinaisons de SNPs

- MLMM: Segura et coll. (Nat Genet, 2012)
- BLMM: Wen (Biostat, 2015)

# Remerciements


- Matthew Stephens
- Xiaoquan Wen, Xiang Zhou et Heejung Shim

```
print(sessionInfo(), locale=FALSE)
```

```
## R version 3.2.2 (2015-08-14)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 14.04.4 LTS
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  me
##
## other attached packages:
## [1] qvalue_2.2.2    knitr_1.12.3    rmarkdown_0.9.2
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.3      digest_0.6.9     plyr_1.8.3
##  [5] gtable_0.1.2     formatR_1.2.1    magrittr_1.5
##  [9] scales_0.3.0     ggplot2_2.0.0    stringi_1.0-1
## [13] splines_3.2.2    tools_3.2.2      stringr_1.0.0
## [17] yaml_2.1.13      colorspace_1.2-6 htmltools_0.3
```