Analyse de données génomiques

Session 3

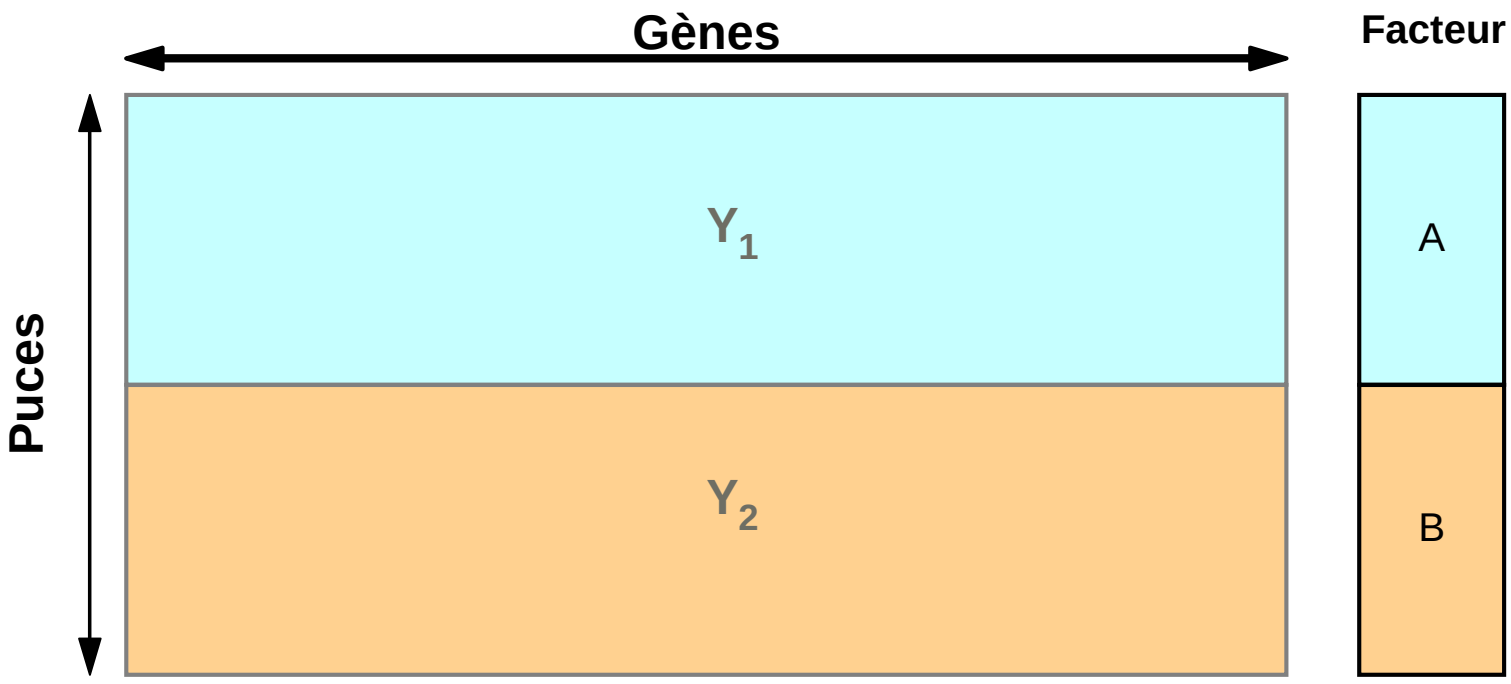
David Causeur

Agrocampus Ouest, IRMAR CNRS UMR 6625

Plan du cours

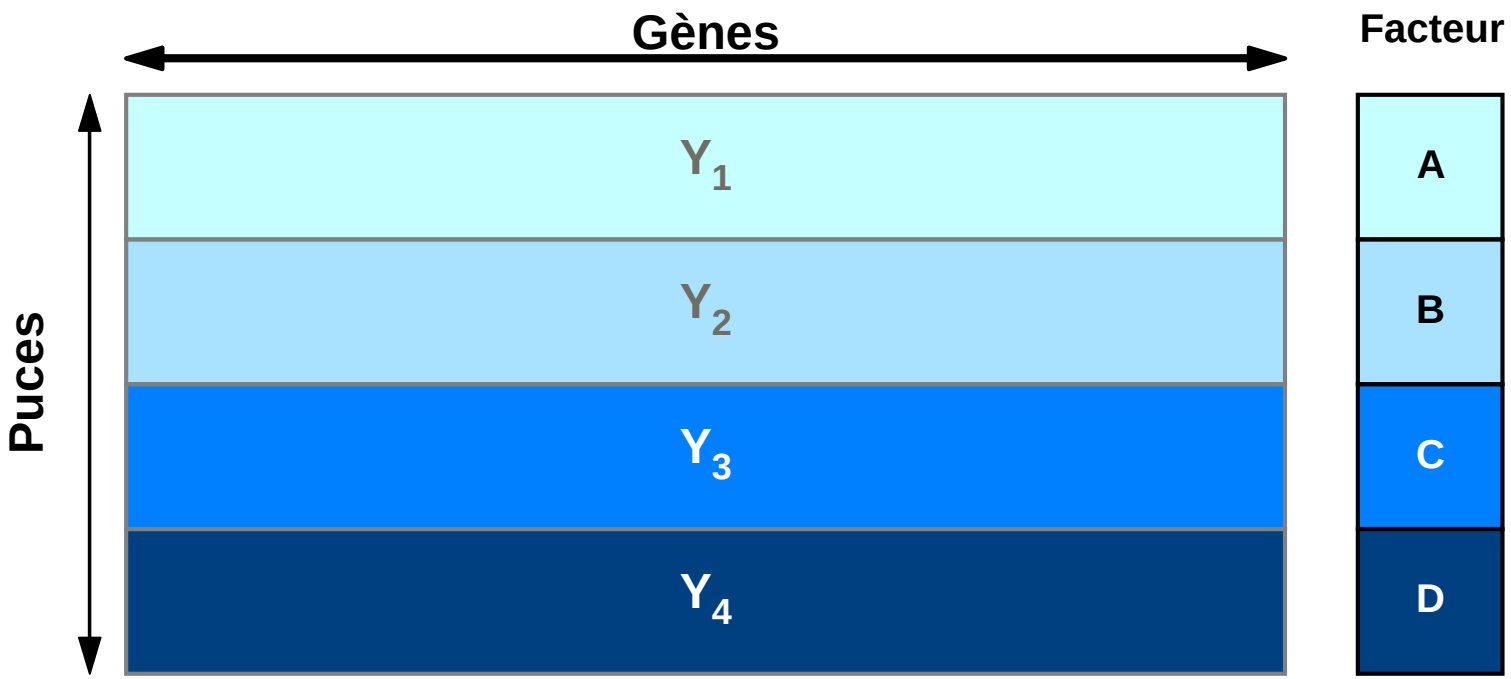
Analyse différentielle

Objectif : identifier les gènes dont l'expression moyenne varie selon les valeurs d'un facteur expérimental



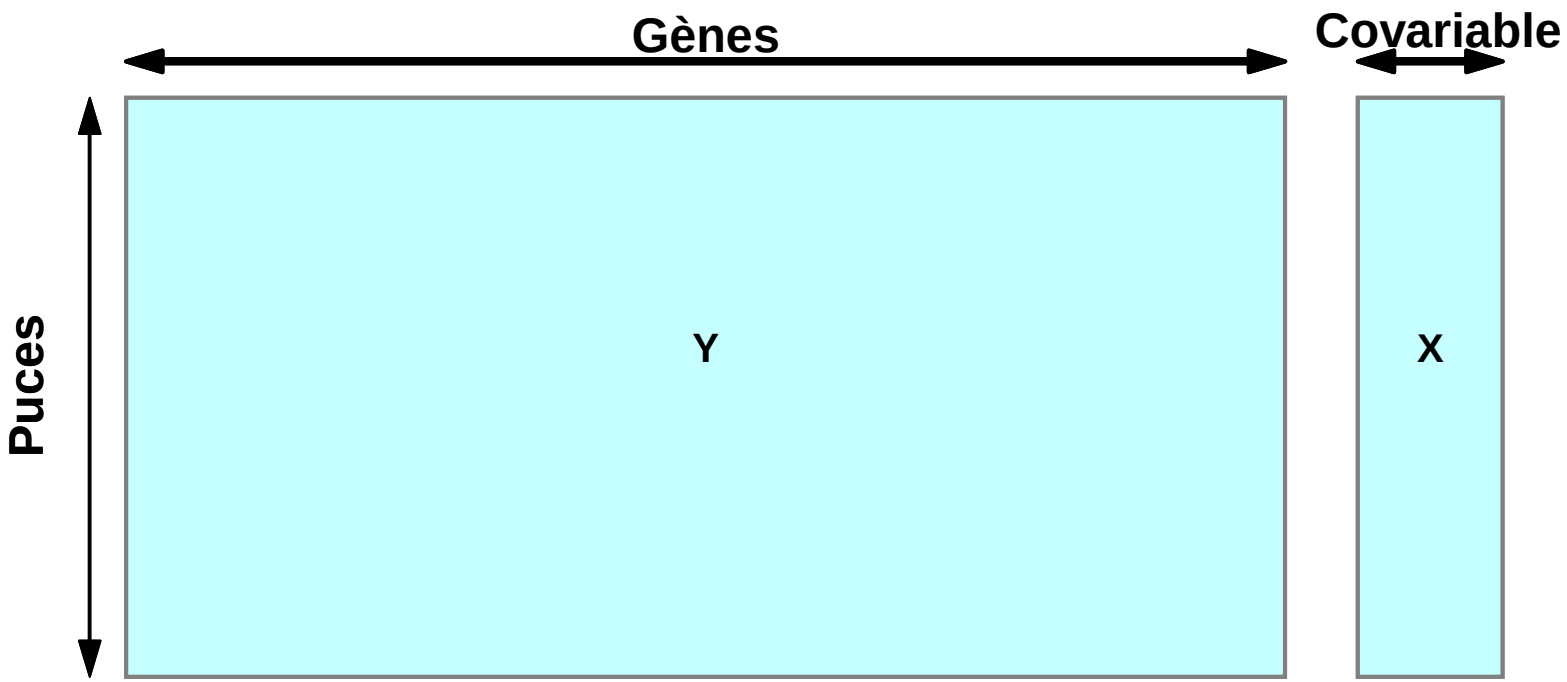
Analyse différentielle

Objectif : identifier les gènes dont l'expression moyenne varie selon les valeurs d'un facteur expérimental



Analyse différentielle

Objectif : identifier les gènes dont l'expression moyenne varie selon les valeurs d'un facteur expérimental



Choix du test

Modèle linéaire : Y expression d'un gène

- Différence entre groupes

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad \text{[effet groupe]}$$

- Co-variation avec une variable continue

$$Y = \beta_0 + \beta_1 x + \varepsilon \qquad \text{[effet linéaire]}$$

Choix du test

Modèle linéaire : Y expression d'un gène

- Différence entre groupes ajustée d'un autre effet

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$ [effet groupe ajusté]

- Co-variation ajustée avec une variable continue

$$Y_{ij} = \beta_0 + \alpha_i + \beta_1 x_{ij} + \varepsilon_{ij}$$ [effet linéaire par groupes]

Choix du test

Modèle linéaire : Y expression d'un gène

- Effet groupe différent selon la modalité d'un autre facteur

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$ [effet d'interaction]

- Co-variation par groupe avec une variable continue

$$Y_{ij} = \beta_0 + \alpha_i + (\beta_1 + \gamma_i) x_{ij} + \varepsilon_{ij}$$ [effet linéaire par groupes]

Plan du cours

Tests multiples

Une collection d'hypotheses nulles $H_0^{(k)},\ k = 1, \ldots, m$

Parmi elles, m_0 "vraies nulles" ... $\mathcal{G}_0 = \{k,\ H_0^{(k)}\ \textit{vraie}\}$

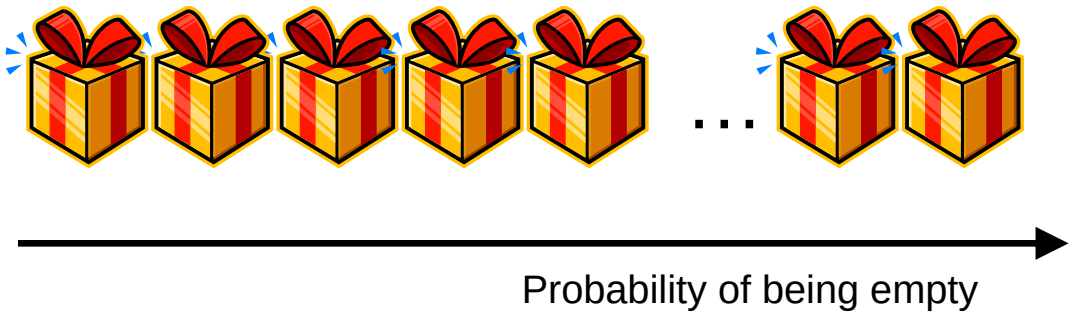
m boxes out of which m_0 are empty

For the kth box, $H^{(k)}_0$: *box is empty*

Tests multiples

Pour le $k$ème test, une p-value $p_k = \mathbb{P}_{H_0^{(k)}}(\text{rejeter } H_0^{(k)})$

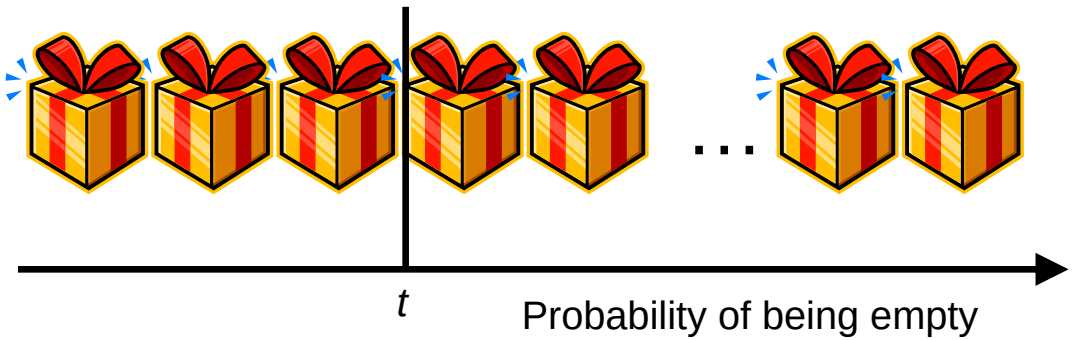
I weight each box … evaluate its probability p of being empty …



Tests multiples

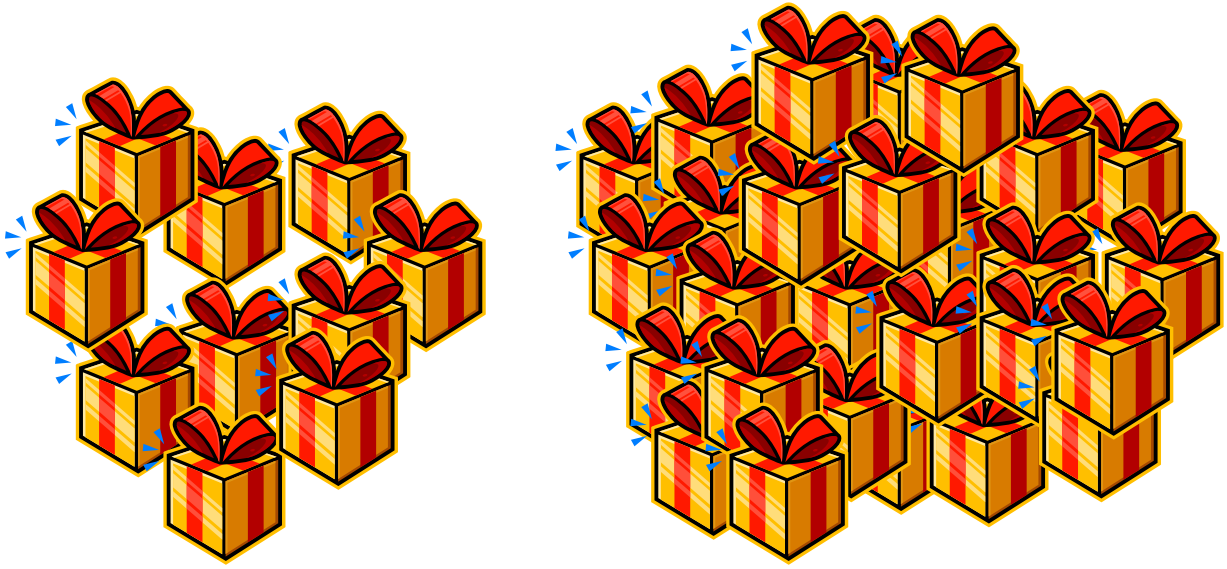
Pour le $k$ème test, une p-value $p_k = \mathbb{P}_{H_0^{(k)}}(\text{rejeter } H_0^{(k)})$

I weight each box … evaluate its probability p of being empty … choose a threshold t



Tests multiples

Pour le $k$ème test, une p-value $p_k = \mathbb{P}_{H_0^{(k)}}(\text{rejeter } H_0^{(k)})$



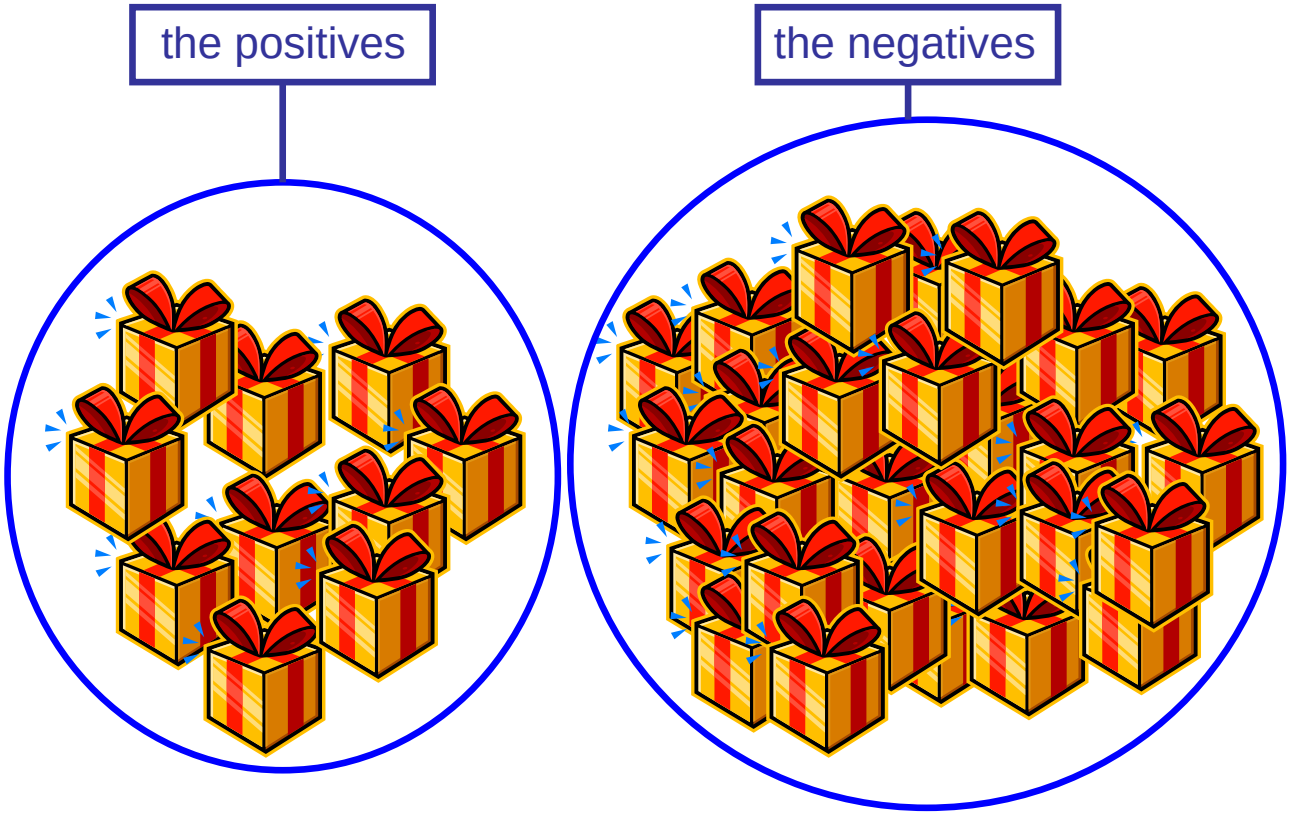
take these ones
$p < t$

…

and declare these empty
$p > t$

Tests multiples

Pour le $k$ème test, une p-value $p_k = \mathbb{P}_{H_0^{(k)}}(\text{rejeter } H_0^{(k)})$



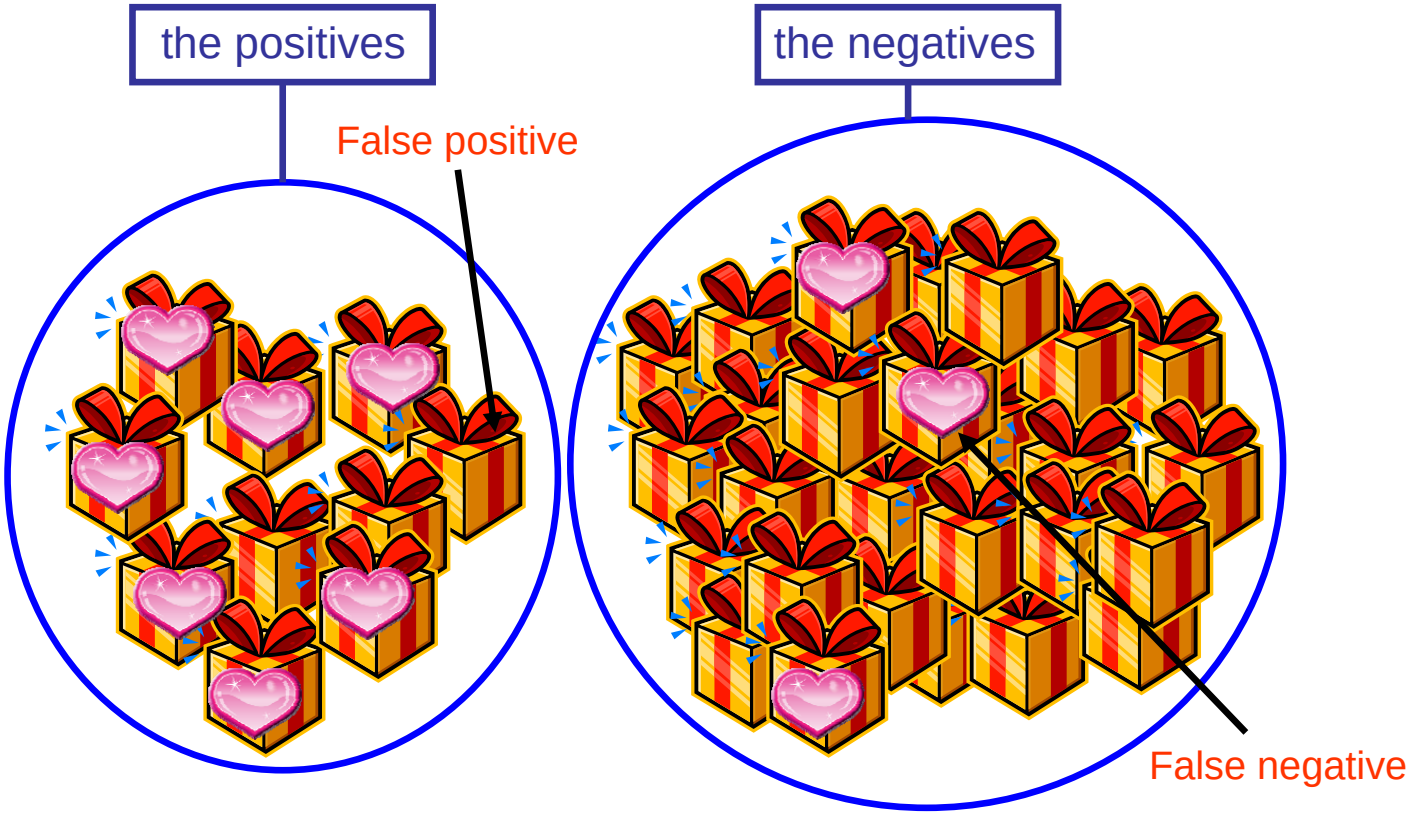
take these ones
$p < t$

…

and declare these empty
$p > t$

Tests multiples

Pour le $k$ème test, une p-value $p_k = \mathbb{P}_{H_0^{(k)}}(\text{rejeter } H_0^{(k)})$



take these ones $p < t$ … and declare these empty $p > t$

Tests multiples

Qu'est-ce qu'une procédure performante ?

- Un nombre contrôlé de faux positifs ; aussi peu de mauvaises surprises que possible ...
- Une grande proportion de positifs parmi les non-nuls ; le plus de découvertes possible ...

Comment construire une procédure performante ?

- Un dispositif expérimental puissant : si le cadeau est bien plus lourd que la boîte ... facile !
- Un bon choix du seuil t de décision !

Contrôle du rique de faux positifs

Pour un seuil t,

- P_t : nombre de gènes positifs (observé)
- FP_t : nombre de gènes faux positifs (non-observé)

t est choisi parmi les p-values :

$$p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(i-1)} \leq t = p_{(i)} \leq p_{(i+1)} \leq \ldots \leq p_{(m)}$$

Si $t = p_{(i)}$, alors $P_t = i$.

Contrôle du rique de faux positifs

Pour un seuil t,

- P_t : nombre de gènes positifs (observé)
- FP_t : nombre de gènes faux positifs (non-observé)

Un objectif possible : garantir que la probabilité qu'il n'y ait aucun faux positif dépasse une valeur fixée (0.95 par exemple)

$$\begin{aligned} \mathbb{P}(FP_t = 0) &\geq 1 - \alpha, \\ \mathbb{P}(FP_t > 0) &\leq \alpha, \end{aligned} \qquad [\text{pour } \alpha = 0.05]$$

Family-Wise Error Rate : $FWER_t = \mathbb{P}(FP_t > 0)$.

Contrôle du rique de faux positifs

Prenons le cas de $m = 10$ gènes

dont $m_0 = 8$ ne sont pas différentiellement exprimés :

$$\mathcal{G}_0 = \{1, 2, \cancel{3}, 4, 5, \cancel{6}, 7, 8, 9, 10\}.$$

$$\begin{aligned} \text{FWER}_t &= \mathbb{P}\Big([p_1 \leq t] \text{ou} [p_2 \leq t] \text{ou} [p_4 \leq t] \text{ou} \ldots \text{ou} [p_{10} \leq t]\Big), \\ &\leq \mathbb{P}[p_1 \leq t] + \mathbb{P}[p_2 \leq t] + \mathbb{P}[p_4 \leq t] + \ldots + \mathbb{P}[p_{10} \leq t], \\ &\leq m_0 t \end{aligned}$$

Si $t = \alpha/m$, alors $\text{FWER}_t \leq \frac{m_0}{m}\alpha \leq \alpha$

Contrôle du rique de faux positifs

Prenons le cas de $m = 10$ gènes

dont $m_0 = 8$ ne sont pas différentiellement exprimés :

$$\mathcal{G}_0 = \{1, 2, \cancel{3}, 4, 5, \cancel{6}, 7, 8, 9, 10\}.$$

$$\begin{aligned} \text{FWER}_t &= \mathbb{P}\Big([p_1 \leq t] \text{ou} [p_2 \leq t] \text{ou} [p_4 \leq t] \text{ou} \ldots \text{ou} [p_{10} \leq t]\Big), \\ &\leq \mathbb{P}[p_1 \leq t] + \mathbb{P}[p_2 \leq t] + \mathbb{P}[p_4 \leq t] + \ldots + \mathbb{P}[p_{10} \leq t], \\ &\leq m_0 t \end{aligned}$$

Si $t = \alpha/m$, alors $\text{FWER}_t \leq \frac{m_0}{m}\alpha \leq \alpha$

p-values ajustées : $p_i \leq \alpha/m \Leftrightarrow \underbrace{mp_i}_{\tilde{p}_i} \leq \alpha$

Contrôle du taux de faux positifs

Pour un seuil t,

- P_t : nombre de gènes positifs (observé)
- FP_t : nombre de gènes faux positifs (non-observé)

Un objectif possible : garantir que le taux de faux positifs ne dépasse pas une valeur fixée (0.05 par exemple)

$$FDR_t = \mathbb{E}\left(\frac{FP_t}{P_t}\right) \quad \leq \quad \alpha, \qquad [\text{pour } \alpha = 0.05]$$

False Discovery Rate : FDR_t.

Contrôle du taux de faux positifs

Méthode de Benjamini-Hochberg (1995)

$$\begin{aligned}\widehat{\mathrm{FDR}}_t &= \frac{\widehat{\mathbb{E}}(\mathrm{FP}_t)}{\mathrm{P}_t}, \\ &= \frac{m_0 t}{\mathrm{P}_t} = \frac{m_0}{m}\frac{mt}{\mathrm{P}_t} \approx \frac{mt}{\mathrm{P}_t}, \text{ si } \pi_0 = \frac{m_0}{m} \approx 1\end{aligned}$$

On choisit t comme la plus grande p-value $p_{(i)}$ telle que

$$\begin{aligned}\widehat{\mathrm{FDR}}_t &\leq \alpha, \text{ avec } t = p_{(i)}, \\ \text{p-values ajustées : } \tilde{p}_{(i)} = \frac{m p_{(i)}}{i} &\leq \alpha.\end{aligned}$$

On garantit ainsi que $\mathrm{FDR}_t \leq \frac{m_0}{m}\alpha \leq \alpha$

Contrôle du taux de faux positifs

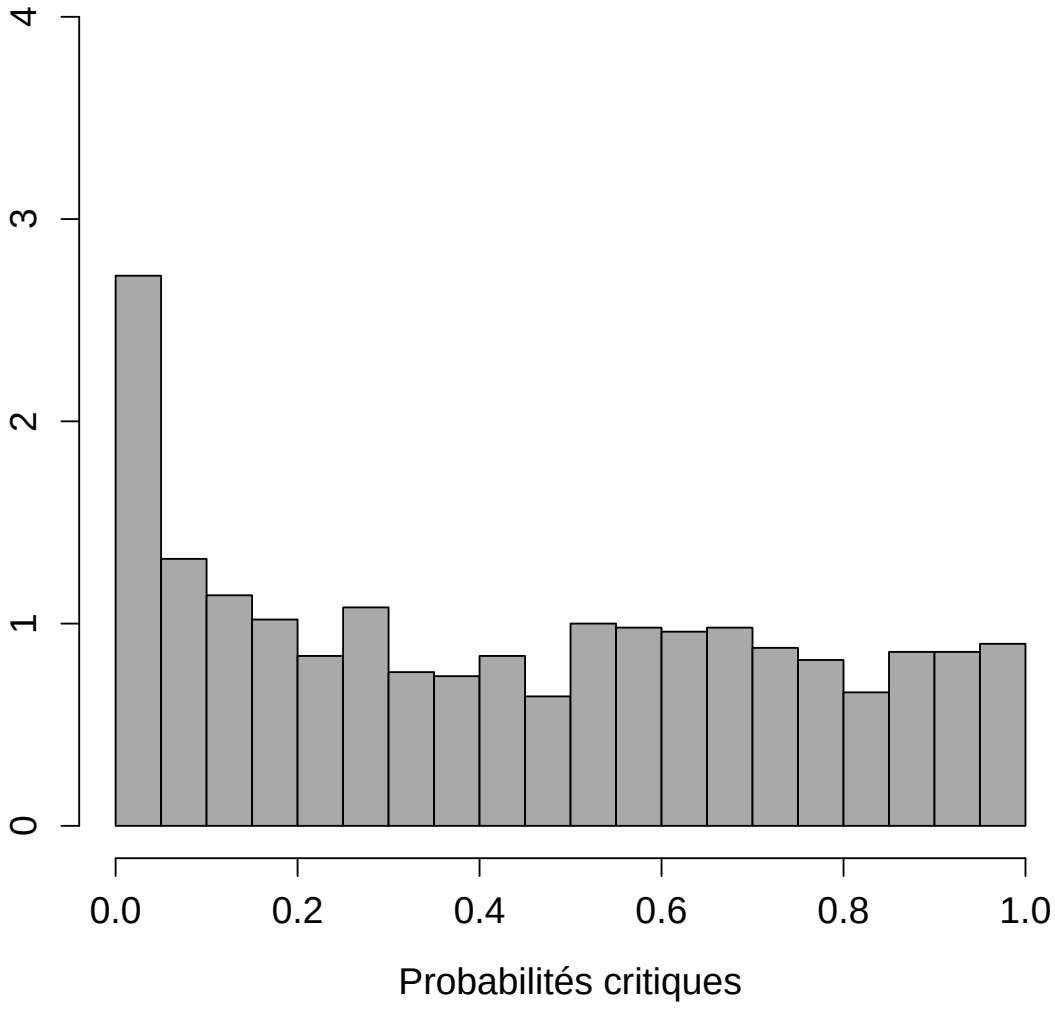
Prenons le cas de $m = 10$ gènes

	Gènes G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
p-val.	0.0001	0.002	0.003	0.01	0.025	0.03	0.07	0.08	0.18	0.75
Bonf.	0.001	0.02	0.03	0.10	0.25	0.30	0.70	0.80	1	1
BH	0.001	0.01	0.01	0.025	0.05	0.05	0.10	0.10	0.20	0.75

Plan du cours

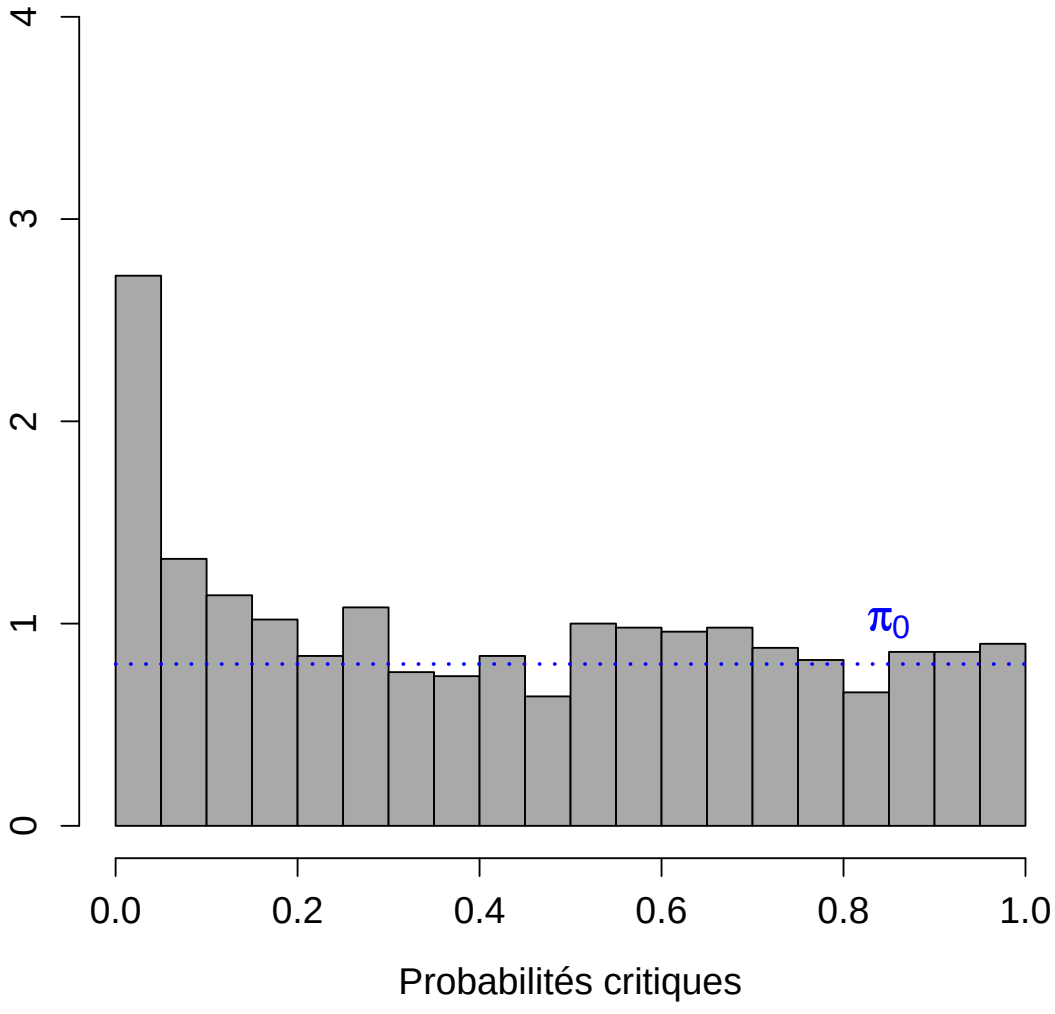
Estimation de π_0

Estimation à partir de la distribution des p-values



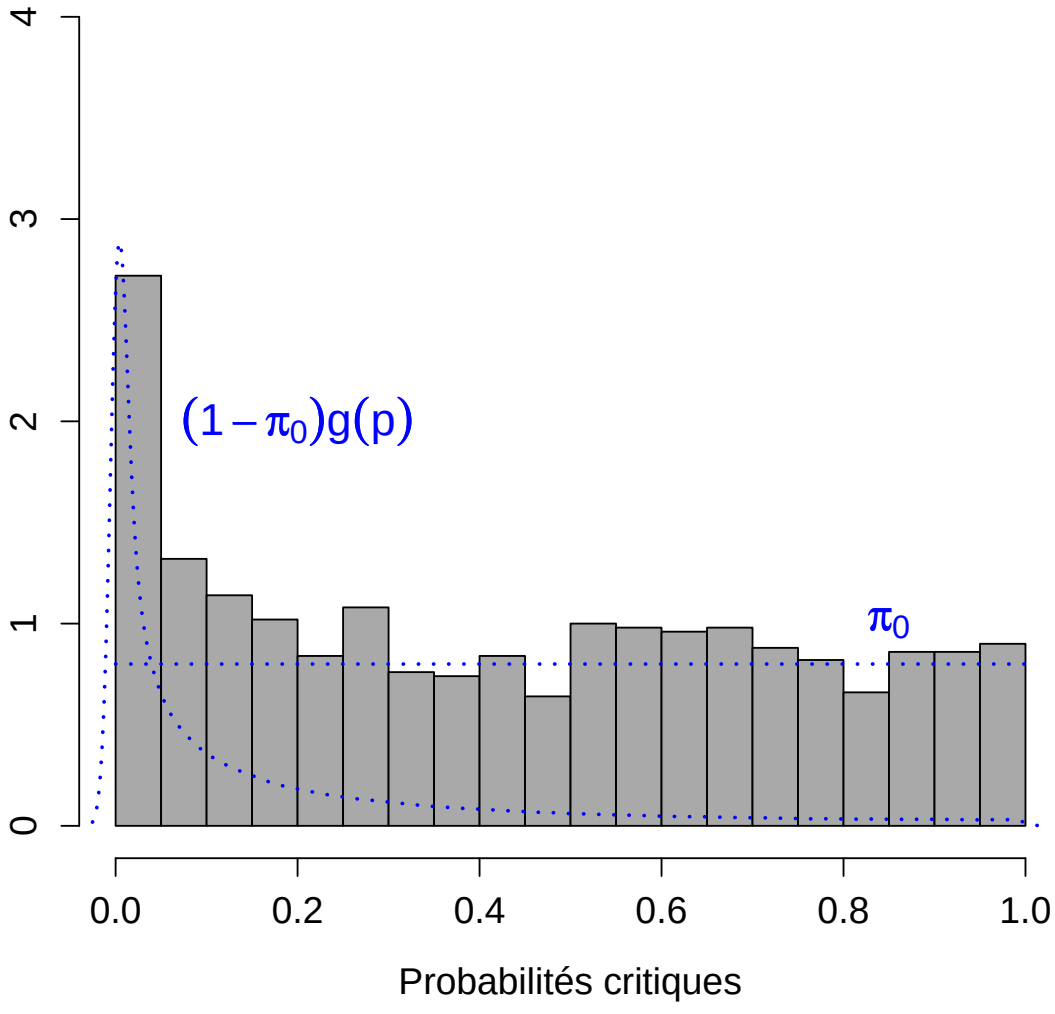
Estimation de π_0

Estimation à partir de la distribution des p-values



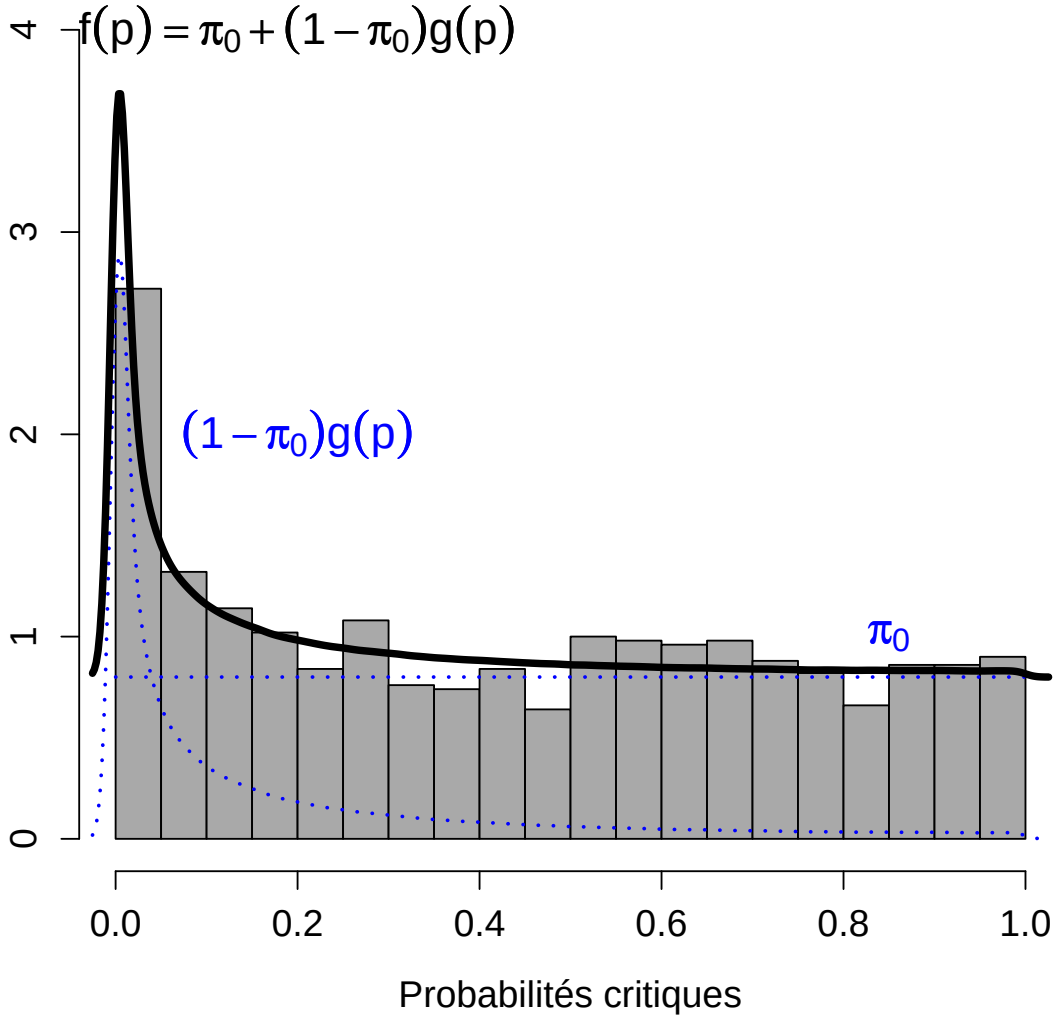
Estimation de π_0

Estimation à partir de la distribution des p-values



Estimation de π_0

Estimation à partir de la distribution des p-values



Estimation de π_0

Un estimateur possible : $\hat{\pi}_0 = \hat{f}(1)$

Amélioration de la procédure de Benjamini-Hochberg :

$$\text{q-value} = \widetilde{\text{FDR}}_t = \hat{\pi}_0 \widehat{\text{FDR}}_t.$$

Si $t = \max\left\{0 \leq t \leq 1,\ \widetilde{\text{FDR}}_t \leq \alpha\right\}$, $\text{FDR}_t \leq \alpha$

Tests modérés

Tests de Fisher (F-tests)

$$\text{Pour le } k\text{ème gène, } F_k = \frac{\text{Variance expliquée pour le gène } k}{\text{Variance résiduelle pour le gène } k},$$

$$= \frac{\text{Variance expliquée pour le gène } k}{s_k^2}.$$

Tests de Fisher modérés

$$\text{Pour le } k\text{ème gène, } F_k = \frac{\text{Variance expliquée pour le gène } k}{s_k^{\star 2}},$$

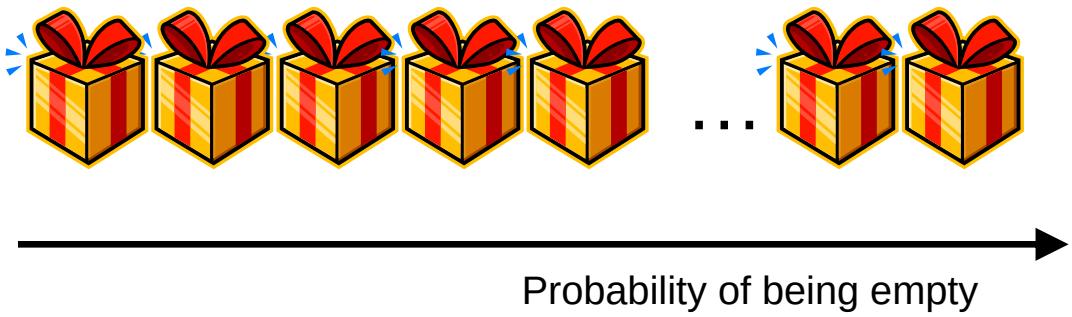
où $s_k^{\star 2} = q_k s_0^2 + p_k s_k^2$.

Estimation des coefficients p_k, q_k par une méthode dite de Bayes empirique.

Hétérogénéité des données d'expression

Une collection d'hypothèses nulles $H_0^{(k)}$

I weight each box … evaluate its probability p of being empty …



Si les boîtes sont de poids différents (certaines en carton, d'autres en métal ...)

... alors le classement ci-dessus n'est plus consistant !

Hétérogénéité des données d'expression

$Y^{(k)}$, expression du $k$ème gène

- Données homogènes

$$Y^{(k)} = \text{signal biologique} + \varepsilon^{(k)}$$

- Données hétérogènes (un facteur d'hétérogénéité)

$$Y^{(k)} = \text{signal biologique} + b_k z + e^{(k)},$$

où z est une composante latente d'hétérogénéité

- Données ajustées de l'hétérogénéité

$$Y^{(k)} - b_k z = \text{signal biologique} + e^{(k)},$$

Hétérogénéité des données d'expression

$Y^{(k)}$, expression du $k$ème gène

- Données homogènes

$$Y^{(k)} = \text{signal biologique} + \varepsilon^{(k)}$$

- Données hétérogènes (q facteurs d'hétérogénéité)

$$Y^{(k)} = \text{signal biologique} + b_{1k}z_1 + \ldots + b_{qk}z_q + e^{(k)},$$

où $z_1, \ldots, z_q$ sont q composantes d'hétérogénéité

- Données ajustées de l'hétérogénéité

$$Y^{(k)} - b_{1k}z_1 - \ldots - b_{qk}z_q = \text{signal biologique} + e^{(k)},$$