## 5. Contrôle du FDR

Considérons une procédure de seuillage

$$R^t(p) = \{ i = 1, \dots, m \quad p_i \leq t(p) \}$$

On peut alors calculer

$$FDR(R^t(p), p) = \mathbb{E}\left( \frac{\#(R^t(p) \cap \mathcal{H}_0(p))}{\#(R^t(p)) \vee 1} \right)$$

$$= \mathbb{E}\left( \frac{\sum_{i \in \mathcal{H}_0(p)} \mathbb{1}(p_i \leq t(p))}{\left(\sum_{i=1}^m \mathbb{1}(p_i \leq t(p))\right) \vee 1} \right)$$

$$= \frac{\alpha}{m} \mathbb{E}\left( \sum_{i \in \mathcal{H}_0(p)} \frac{\mathbb{1}(p_i \leq t(p))}{\left(\frac{\alpha}{m} \sum_{i=1}^m \mathbb{1}(p_i \leq t(p))\right) \vee \frac{\alpha}{m}} \right) \quad (*)$$

On reconnait la fdr empirique des p-values.

Notons $\hat{G}_m(p, x) = \frac{1}{m} \sum_{i=1}^m \mathbb{1}(p_i \leq x)$.

⚠ Avec les mains : si $\hat{G}_m$ était déterministe, de même que $t(p)$, on aimerait choisir $t$ tel que :

$$t \leq \alpha \hat{G}_m(p, t) \iff t \vee \frac{\alpha}{m} \leq \alpha \hat{G}_m(p, t) \vee \frac{\alpha}{m}$$

car dans ce cas on aurait.

$$\frac{\alpha}{m} \mathbb{E}\left( \sum_{i \in \mathcal{H}_0(p)} \frac{\mathbb{1}(p_i \leq t(p))}{\alpha \hat{G}_m(p, t(p)) \vee \frac{\alpha}{m}} \right)$$

$$\leq \frac{\alpha}{m} \frac{1}{t(p) \vee \frac{\alpha}{m}} \sum_{i \in \mathcal{H}_0(p)} \underbrace{\mathbb{E}(\mathbb{1}(p_i \leq t(p)))}_{= \text{"}\mathbb{P}(p_i \leq t(p))\text{"} \leq t(p)}$$

par propriété des p-values

$$\leq \frac{\alpha}{m} \frac{1}{t(p) \vee \frac{\alpha}{m}} \sum_{i \in \mathcal{H}_0(p)} t(p) \leq \frac{\alpha}{m} \frac{1}{t(p) \vee \frac{\alpha}{m}} t(p) m_0(p)$$

$$\leq \alpha \frac{m_0(p)}{m} \leq \alpha .$$

On introduit des seuils dits "auto consistants"

$$t \in \mathcal{I}(p) \qquad \mathcal{I}(p) = \{ u \in [0,1], \hat{G}_m(p, u) \geq \frac{u}{\alpha} \}$$

On les dit auto consistants car si $t \in \mathcal{I}(p)$, on a

$$\hat{G}_m(p, t) \geq \frac{t}{\alpha} \iff \#(p_i \leq t) \geq \frac{t\, m}{\alpha} \quad (**)$$

Donc pour un seuil $t$, on écrit : si $t \in \mathcal{I}(p)$

$$R^t(p) = \{ i = 1, \dots, m . \ p_i \leq t \} \subset \{ i = 1, \dots, m, \ p_i \leq \frac{\alpha}{m} \#(R^t(p)) \}.$$

Un seuil auto-consistant sera en pratique très simple à utiliser car il ne dépend que de $\alpha$ et des p-valeurs $p$.

Si on choisit $t(p) \in \mathcal{T}(p)$, on peut écrire.

$$FDR(R^t(p), P) \leq \frac{\alpha}{m} \sum_{i \in \mathcal{H}_0(P)} \mathbb{E}\left(\frac{\mathbb{1}(p_i \leq t(p))}{t(p) \vee \frac{\alpha}{m}}\right) \leq \frac{\alpha}{m} \sum_{i \in \mathcal{H}_0(P)} \mathbb{E}\left(\frac{\mathbb{1}(p_i \leq t(p))}{t(p)}\right)$$

on a utilisé (*) et (**)

avec la convention $\frac{0}{0} = 0$

**Lemme** (H)

Soit $U \geq 0$ v.a. telle que $\mathbb{P}(U \leq u) \leq u$ $\forall u \in [0,1]$ et $g$ une fonction de $\mathbb{R}_+$ dans $\mathbb{R}_+$ décroissante. On pose $V = g(U)$ alors $\mathbb{E}\left(\frac{\mathbb{1}(U \leq V)}{V}\right) \leq 1$

Preuve: Soient $\mathcal{U} = \{u \geq 0, g(u) \geq u\}$ et $u^* = \sup \mathcal{U}$.

On pose $C^* = \inf_{u \in \mathcal{U}} g(u)$.

Pour tout $u$ dans $\mathcal{U}$, on a $u \leq u^* \leq g(u^*)$. Comme $g$ est décroissante et que $u^* \geq u$, on a $g(u^*) \leq g(u)$ (c'est vrai pour tout $u \in \mathcal{U}$)

donc $g(u^*) \leq C^*$ (on peut choisir une suite $(u_n)$ dans $\mathcal{U}$ croissante vers $u^*$ $g(u_n) \searrow C^*$ )

On en déduit que $C^* \geq u^*$.

On a donc $C^* = 0 \Rightarrow u^* = 0$

sinon.

$$\mathbb{E}\left(\frac{\mathbb{1}(U \leq g(U))}{g(U)}\right) \leq \frac{\mathbb{E}(\mathbb{1}(U \leq g(U)))}{C^*} = \frac{\mathbb{P}(U \leq g(U))}{C^*}$$

$$\leq \frac{\mathbb{P}(U \leq u^*)}{C^*} \leq \frac{u^*}{C^*} \leq 1$$

via l'hyp. (H)

On se restreint maintenant à la famille des lois suivantes

$$\mathcal{P}^I = \{P \in \mathcal{P} \quad (p_i(X))_{i \in \mathcal{H}_0(P)} \text{ indépendantes}, (p_i(X))_{i \in \mathcal{H}_1(P)} \text{ indépendantes de } (p_i(X))_{i \in \mathcal{H}_0(P)}\}.$$

C'est une famille très restrictive qui demande l'indépendance des p-values. Cela peut être amélioré. On trouve par exemple dans des papiers récents $P$ loi gaussienne multivariée avec covariances $\geq 0$ (cf notamment Benjamini & Yekutieli, Annals of Stat.) 2001

On peut dans cette famille et grâce au lemme montrer le théorème suivant.

**Théorème**
Soit $R^t$ une procédure de tests multiples basée sur une fonction de seuil $t(.)$ vérifiant :
(*) $t$ est auto-consistante
(**) $t(q) \leq t(q')$ $\forall q, q' \in [0,1]^m$ vérifiant $q_i' \leq q_i$ $\forall i = 1, \dots, m$
(i.e $t$ est "décroissante" par rapport à toutes les p-valeurs)
alors $FDR(R^t(p), p) \leq \frac{\alpha m_0(P)}{m} \leq \alpha$ $\forall P \in \mathcal{P}^I$

Preuve : Soit $i \in \mathcal{H}_0(P)$, on reprend le calcul du haut de la page 8 et on s'intéresse au terme.

$$\mathbb{E}\left(\frac{\mathbb{1}(p_i \leq t(p))}{t(p)}\right)$$

L'astuce est de conditionner par rapport à $p_{(-i)} = (p_1, \dots, p_{i-1}, p_{i+1}, p_m)$ la fonction $t^{(i)} : p_i \to t(p_i ; p_{(-i)}) = t^{(i)}(p_i)$ (on fixe $p_{(-i)}$) est décroissante en $p_i$ par l'hypothèse (**) du théorème.
On peut donc appliquer le lemme avec $g = t^{(i)}$ et $U = p_i$

$$\mathbb{E}\left(\frac{\mathbb{1}(p_i \leq t(p))}{t(p)}\right) = \mathbb{E}\left(\mathbb{E}\left(\frac{\mathbb{1}(p_i \leq t^{(i)}(p_i))}{t^{(i)}(p_i)} \,\Big|\, p_{(-i)}\right)\right)$$

L'hypothèse d'indépendance sert ici à rendre la fonction $t^{(-i)}$ déterministe après conditionnement.

$\leq 1$ car $p_i \mapsto t^{(i)}(p_i)$ est décroissante et $\mathbb{P}(p_i \leq u) \leq u$ ce qui prouve le théorème. □

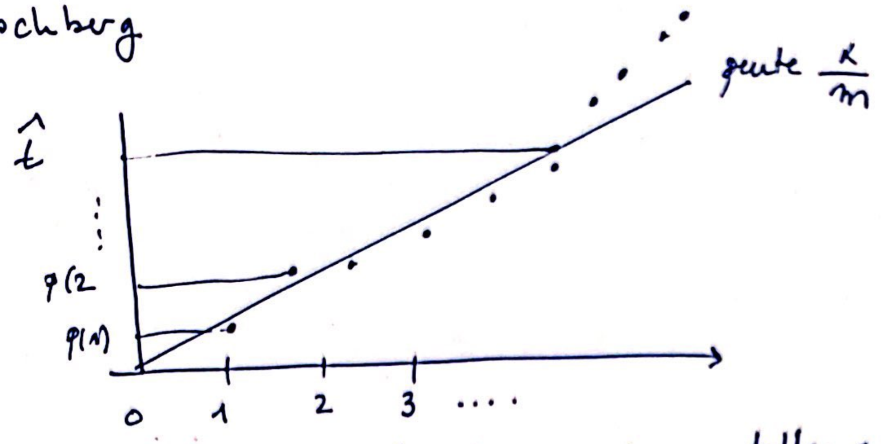
## 6. La procédure de Benjamini-Hochberg (1994)

Le théorème précédent montre que, sous une hypothèse sur les p-valeurs, toute procédure de test basée sur une fonction de seuil décroissante et auto-consistante contrôle le FDR. Il est donc naturel de considérer le seuil suivant (pour maximiser $R^t$) :

$$\begin{aligned}\hat{t}(p) &= \max \mathcal{T}(p)\\ &= \max\left\{u \in \left\{\frac{\alpha k}{m}, k = 0, \dots, m\right\}, \hat{G}_m(p, u) \geq \frac{u}{\alpha}\right\}\\ &= \frac{\alpha}{m} \max\left\{k = 0, \dots, m : p_{(k)} \leq \frac{\alpha k}{m}\right\}\end{aligned}$$

où $p_{(1)} \leq \dots \leq p_{(m)}$.

Ce choix est appelé linear-step-up ou seuillage de Benjamini-Hochberg




et la procédure rejette les p-valeurs telles que $p_i \leq \hat{t}$.

Note : il existe des démonstrations plus élégantes, voir notamment statweb.stanford.edu/~ckirby/brad/LSI/chapter4.pdf avec un argument martingale et un théorème d'arrêt optionnel.