Section 11

Tests multiples

Séance du 9 avril 2027

Lectures : Farcomeni (2010), ch. 3 de *Medical Biostatistics for Complex Diseases* ; Dudoit et van der Laan (2008), ch. 1 à 4 (survol)

Julien Miron – notes rédigées à partir du matériel de Lajmi Lakhal-Chaieb et Alexandre Bureau

Objectifs de la séance

À la fin de cette séance, vous devriez être en mesure de :

1. décrire les issues possibles d'une famille de m tests et définir les principaux taux d'erreur de 1re espèce (FWER, gFWER, FDP, FDR, FDX/TPPFP) ainsi que les mesures de puissance associées (FNR, TDR) ;
2. démontrer les relations entre ces taux (en particulier FDR $\leq$ FWER, avec égalité sous l'hypothèse nulle complète) et distinguer contrôle faible et contrôle fort ;
3. appliquer à la main et avec `multtest` les procédures de Bonferroni, Šidák, Holm, Hochberg, Benjamini–Hochberg (BH) et Benjamini–Yekutieli (BY), et calculer des valeurs p ajustées ;
4. expliquer les procédures minP et maxT et le rôle de la loi nulle conjointe (permutation ou bootstrap) ;
5. choisir entre FWER et FDR selon le contexte et calculer le nombre attendu de rejets et de faux rejets.

1 Mise en contexte : des milliers de tests à la fois

Les études biomédicales de « haut débit » produisent couramment des milliers de variables pour chaque sujet : niveaux d'expression de gènes mesurés par micropuces ou par séquençage d'ARN, génotypes de centaines de milliers de polymorphismes, abondances de protéines ou de métabolites. La question type est descriptive et exploratoire : *lesquels* de ces gènes sont exprimés différemment chez les patients atteints d'une leucémie myéloïde aiguë (AML) et chez ceux qui sont atteints d'une leucémie lymphoblastique aiguë (ALL) ? On teste alors, pour chaque gène j, une hypothèse nulle $H_0(j)$ d'absence de différence de moyenne de log-expression entre les deux groupes. Le nombre d'hypothèses peut atteindre $m = 10^4$ à 10^6 (Farcomeni, § 3.1).

Le problème est simple à énoncer. Si chaque hypothèse est testée au seuil α sans correction, chaque hypothèse nulle vraie a une probabilité α d'être rejetée. Si m_0 hypothèses nulles sont vraies, le nombre V de faux rejets vérifie, quelle que soit la dépendance entre les tests,

$$\mathrm{E}[V] \;=\; \sum_{j:\, H_0(j) \text{ vraie}} \mathbb{P}\{H_0(j) \text{ rejetée}\} \;=\; m_0\,\alpha.$$

Le nombre attendu de faux positifs est donc *proportionnel au nombre d'hypothèses nulles vraies*. Avec $m = 10\,000$ gènes dont $m_0 = 9\,000$ ne sont pas différentiellement exprimés, on s'attend à 450 faux positifs au seuil 5 % ; ces faux positifs représentent 4,5 % de toutes les hypothèses testées, mais ils peuvent constituer une part importante, voire la majorité, de la liste des gènes « significatifs ». Si en plus les tests sont indépendants, la probabilité de commettre au moins une erreur vaut $1-(1-\alpha)^{m_0}$, qui est pratiquement égale à 1 dès que m_0 dépasse quelques centaines.

Il faut donc choisir un *critère* global d'erreur pour la famille de tests, puis une *procédure* qui le maintient sous un niveau choisi ; ce choix dépend de l'usage qu'on fera des hypothèses rejetées.

Grille méthodologique

1. **Objet d'intérêt** : l'ensemble $\mathcal{H}_1 = \{j : H_0(j) \text{ fausse}\}$ des hypothèses nulles fausses (p. ex. les gènes différentiellement exprimés), qu'on cherche à estimer par l'ensemble $\mathcal{R}$ des hypothèses rejetées.
2. **Nuisance et sources de biais** : la proportion $\pi_0 = m_0/m$ d'hypothèses nulles vraies, la structure de dépendance entre les statistiques de test, les lois marginales des statistiques (qui peuvent différer d'un test à l'autre) ; biais de sélection si on ne retient que les résultats « significatifs ».
3. **Approche** : définir un taux d'erreur de 1re espèce pour la famille (FWER, FDR, …) et le contrôler, de façon marginale (à partir des seules valeurs p) ou conjointe (à partir de la loi nulle conjointe des statistiques).
4. **Critère** : rejeter le plus d'hypothèses possible sous la contrainte « taux d'erreur $\leq \alpha$ », ce qui se traduit par des valeurs p ajustées.
5. **Calcul** : seuils explicites (Bonferroni, Holm, BH, …) ou algorithmes de ré-échantillonnage (Westfall–Young, bootstrap de Dudoit–van der Laan).
6. **Inférence** : contrôle exact (en échantillon fini) ou asymptotique selon la façon d'obtenir la loi nulle des statistiques (permutation ou bootstrap) et selon la dépendance.

2 Cadre et notation

On dispose d'un échantillon de taille n et de m hypothèses nulles $H_0(1),\ldots,H_0(m)$, chacune portant sur un paramètre $\psi(j)$ de la loi des données, typiquement $H_0(j) : \psi(j) = \psi_0(j)$ ou $H_0(j) : \psi(j) \leq \psi_0(j)$. On note $T_n(j)$ la statistique de test et p_j la valeur p (« brute » ou non ajustée) associée. Soit $\mathcal{H}_0 \subseteq \{1,\ldots,m\}$ l'ensemble (inconnu) des hypothèses nulles vraies, $m_0 = |\mathcal{H}_0|$, $m_1 = m - m_0$ et $\pi_0 = m_0/m$. Une procédure de tests multiples produit un ensemble d'hypothèses rejetées $\mathcal{R}$, de cardinalité $R = |\mathcal{R}|$. L'hypothèse nulle complète est $H_0^C = \bigcap_{j=1}^m H_0(j)$, c'est-à-dire $m_0 = m$.

On suppose que, pour $j \in \mathcal{H}_0$, la valeur p est *valide* : $\mathbb{P}(p_j \leq u) \leq u$ pour tout $u \in [0,1]$; pour un test asymptotique (Wald, rapport de vraisemblance), ce n'est vrai qu'à la limite.

Le tableau 1 résume les issues de la famille de tests. Seules m et R sont observées ; toutes les autres quantités sont des variables aléatoires non observables.

TABLE 1 – Issues d'une famille de m tests. Notation de ces notes ; entre crochets, la notation de Farcomeni (2010), où $N_{a|b}$ compte les hypothèses dont la décision est a (1 = rejet) et la vérité est b (1 = H_0 fausse). Dudoit et van der Laan (2008) écrivent V_n et R_n pour souligner la dépendance envers n.

	Non rejetée	Rejetée	Total
H_0 vraie	U $[N_{0\|0}]$	V $[N_{1\|0}]$	m_0
H_0 fausse	W $[N_{0\|1}]$	S $[N_{1\|1}]$	m_1
Total	$m - R$	R $[R_n]$	m

Ainsi $V = |\mathcal{R} \cap \mathcal{H}_0|$ est le nombre de *faux positifs* (erreurs de 1re espèce), W le nombre de *faux négatifs* (erreurs de 2e espèce) et $S = R - V$ le nombre de *vraies découvertes*. Le choix de W (plutôt que le T de Benjamini et Hochberg, 1995) évite la confusion avec les statistiques de test.

3 Taux d'erreur de 1re espèce

3.1 Définitions

Un taux d'erreur de 1re espèce pour une famille de tests est un paramètre de la loi de V ou de la loi conjointe de (V,R). Dudoit et van der Laan insistent sur ce point : un taux d'erreur est une fonctionnelle $\theta(F_{V_n,R_n})$, et on peut ainsi traiter tous les critères dans un même cadre.

Définition 3.1 (Taux d'erreur usuels)**.** Soit $\text{FDP} = V/R$ si $R > 0$ et $\text{FDP} = 0$ si $R = 0$ (proportion de fausses découvertes, *false discovery proportion*). On définit :

- le taux d'erreur par famille (*family-wise error rate*) : $\text{FWER} = \mathbb{P}(V > 0) = 1 - F_V(0)$;
- le FWER généralisé, pour un entier $k \geq 0$: $\text{gFWER}(k) = \mathbb{P}(V > k)$; on tolère k faux positifs et $\text{gFWER}(0) = \text{FWER}$;
- le taux de fausses découvertes (*false discovery rate*) : $\text{FDR} = \text{E}[\text{FDP}] = \text{E}\left[\frac{V}{R}\,\mathbf{1}\{R > 0\}\right]$;
- pour $q \in (0,1)$, la probabilité de dépassement de la proportion de fausses découvertes : $\text{FDX}(q) = \text{TPPFP}(q) = \mathbb{P}(\text{FDP} > q)$ (*false discovery exceedance* chez Farcomeni ; *tail probability for the proportion of false positives* chez Dudoit et van der Laan).

Remarque 3.2. Farcomeni définit le k-FWER comme $\mathbb{P}(V \geq k)$; son k-FWER correspond donc au $\text{gFWER}(k-1)$ de Dudoit et van der Laan. Vérifiez toujours la convention utilisée par le logiciel (la fonction `fwer2gfwer` de `multtest` suit Dudoit et van der Laan).

Le FWER et le gFWER portent sur le *nombre* de faux positifs, alors que le FDR et le FDX portent sur leur *proportion* parmi les rejets. C'est la différence essentielle : quand R est grand, quelques faux positifs ne coûtent presque rien en proportion. Le FDR contrôle la moyenne de la FDP ; le FDX contrôle la queue de sa loi, ce qui est plus exigeant quand la FDP est très variable (notamment sous forte dépendance).

Du côté de la puissance, on utilise par analogie :

— le taux de faux négatifs (*false non-discovery rate*) : $\mathrm{FNR} = \mathrm{E}\left[\dfrac{W}{m-R}\,\mathbf{1}\{R<m\}\right]$, proportion attendue d'hypothèses fausses parmi celles qui ne sont pas rejetées ;

— le taux de vraies découvertes (Dudoit et van der Laan, p. 23) : $\mathrm{TDR} = \mathrm{E}\left[\dfrac{R-V}{R}\,\mathbf{1}\{R>0\}\right]$;

— la puissance moyenne $\mathrm{E}[S]/m_1$ (proportion attendue des hypothèses fausses qui sont détectées), la mesure qu'on rapporte le plus souvent en simulation.

3.2 Relations entre les taux

Proposition 3.3. *Pour toute procédure et toute loi des données :*

1. $\mathrm{FDR} \le \mathrm{FWER} \le \mathrm{E}[V]$ *;*
2. *sous l'hypothèse nulle complète* H_0^C *(*$m_0 = m$*),* $\mathrm{FDR} = \mathrm{FWER}$ *;*
3. $\mathrm{TDR} = \mathbb{P}(R>0) - \mathrm{FDR}$ *;*
4. *pour tout* $q \in (0,1)$*,* $\mathrm{FDX}(q) \le \mathrm{FWER}$ *et* $\mathrm{FDR} \le q + (1-q)\,\mathrm{FDX}(q)$*.*

Démonstration. (1) Si $V = 0$, alors $\mathrm{FDP} = 0$. Si $V \ge 1$, alors $R \ge V \ge 1$ et $\mathrm{FDP} = V/R \le 1$. Dans tous les cas, $\mathrm{FDP} \le \mathbf{1}\{V>0\}$; en prenant l'espérance, $\mathrm{FDR} \le \mathbb{P}(V>0) = \mathrm{FWER}$. De plus, comme V est un entier non négatif, $\mathbf{1}\{V>0\} \le V$, d'où $\mathrm{FWER} \le \mathrm{E}[V]$.

(2) Si toutes les hypothèses nulles sont vraies, tout rejet est un faux rejet : $V = R$. Alors $\mathrm{FDP} = 1$ si $R > 0$ et 0 sinon, c'est-à-dire $\mathrm{FDP} = \mathbf{1}\{V>0\}$, d'où l'égalité.

(3) $\dfrac{R-V}{R}\mathbf{1}\{R>0\} = \mathbf{1}\{R>0\} - \mathrm{FDP}$; on prend l'espérance.

(4) $\{\mathrm{FDP} > q\} \subseteq \{V>0\}$. Pour la seconde inégalité, $\mathrm{FDP} \le q$ sur $\{\mathrm{FDP} \le q\}$ et $\mathrm{FDP} \le 1$ ailleurs, donc $\mathrm{E}[\mathrm{FDP}] \le q\,\mathbb{P}(\mathrm{FDP} \le q) + \mathbb{P}(\mathrm{FDP} > q) = q + (1-q)\,\mathrm{FDX}(q)$. □

La conséquence pratique de (1) est capitale : toute procédure qui contrôle le FWER au niveau α contrôle aussi le FDR au niveau α, mais l'inverse est faux. Contrôler le FDR est une contrainte plus faible, qui permet de rejeter plus d'hypothèses, donc d'avoir plus de puissance. Le point (2) montre que l'écart entre les deux critères vient des hypothèses fausses : sous H_0^C, il n'y en a aucun. Le gain du FDR est d'autant plus grand que m_1 et la puissance individuelle sont grands, car les vraies découvertes « diluent » les faux positifs.

Remarque 3.4 (Le FWER croît avec m). À seuil individuel fixe, le FWER croît avec le nombre d'hypothèses testées (sous indépendance, $1-(1-\alpha)^{m_0}$). Pour le maintenir à α, il faut des seuils individuels de l'ordre de α/m, et la puissance par hypothèse s'effondre quand m est grand (Farcomeni, § 3.3 ; Dudoit et van der Laan, p. 145). À l'inverse, on peut garder constante la proportion de faux positifs parmi les rejets, ce que fait le contrôle du FDR.

3.3 Contrôle faible et contrôle fort

Définition 3.5. Une procédure contrôle un taux d'erreur θ au niveau α

— *faiblement* si $\theta \leq \alpha$ sous l'hypothèse nulle complète H_0^C seulement ;

— *fortement* si $\theta \leq \alpha$ quelle que soit la configuration des hypothèses vraies et fausses, c'est-à-dire pour tout $\mathcal{H}_0 \subseteq \{1,\ldots,m\}$.

Le contrôle est *exact* s'il vaut pour tout n, *asymptotique* si $\limsup_{n\to\infty} \theta_n \leq \alpha$.

Le contrôle faible ne suffit pas en pratique : dès qu'on rejette une hypothèse, on admet que H_0^C est fausse, et la garantie disparaît justement pour les hypothèses qu'on continue d'examiner. Par la proposition 3.3 (2), contrôler faiblement le FDR équivaut à contrôler faiblement le FWER. Toutes les procédures présentées plus loin offrent un contrôle fort, sous des hypothèses qui seront précisées.

À retenir

— Sans correction, $\mathrm{E}[V] = m_0\alpha$: le nombre de faux positifs croît avec le nombre d'hypothèses nulles vraies.

— $\mathrm{FWER} = \mathbb{P}(V > 0)$; $\mathrm{FDR} = \mathrm{E}[V/R]$ (avec $0/0 = 0$) ; $\mathrm{FDX}(q) = \mathbb{P}(V/R > q)$.

— $\mathrm{FDR} \leq \mathrm{FWER}$, avec égalité sous H_0^C : contrôler le FWER implique contrôler le FDR ; le FDR est moins exigeant et donne plus de puissance.

— On exige un contrôle *fort*, valide quelle que soit la configuration des hypothèses vraies.

4 Valeurs p ajustées et types de procédures

4.1 Définition générale

Une procédure de tests multiples (PTM) au niveau α définit, pour chaque hypothèse, une région de rejet $\mathcal{C}_n(j;\alpha)$: on rejette $H_0(j)$ si $T_n(j) \in \mathcal{C}_n(j;\alpha)$. Les régions sont emboîtées en α : si on rejette au niveau α, on rejette à tout niveau $\alpha' > \alpha$.

Définition 4.1 (Valeur p ajustée)**.** La valeur p ajustée de l'hypothèse j pour une PTM donnée est le plus petit niveau auquel cette procédure rejette $H_0(j)$:

$$\tilde{p}_j = \inf\{\alpha \in [0,1] : H_0(j) \text{ est rejetée par la PTM au niveau } \alpha\} = \inf\{\alpha \in [0,1] : T_n(j) \in \mathcal{C}_n(j;\alpha)\}.$$

On rejette donc $H_0(j)$ au niveau α si et seulement si $\tilde{p}_j \leq \alpha$ (Farcomeni, § 3.4.1 ; Dudoit et van der Laan, p. 32). L'intérêt est double : (i) comme une valeur p ordinaire, $\tilde{p}_j$ ne dépend pas d'un niveau choisi d'avance, de sorte que le lecteur peut appliquer son propre niveau ; (ii) les valeurs p ajustées de procédures différentes, voire de critères différents (FWER ou FDR), se comparent directement. Une valeur p ajustée pour le FDR est parfois appelée « valeur q ».

4.2 Trois architectures

Soit $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ les valeurs p ordonnées et $H_{(j)}$ l'hypothèse correspondante. On distingue :

— les procédures **à une étape** (*single-step*) : un même seuil c pour toutes les hypothèses ; on rejette $H_0(j)$ si $p_j \le c$;

— les procédures **descendantes** (*step-down*) : on part de la plus petite valeur p et on « descend » la liste ; avec des seuils croissants $\alpha_1 \le \cdots \le \alpha_m$, on rejette $H_{(1)},\ldots,H_{(k^*-1)}$ où $k^* = \min\{j : p_{(j)} > \alpha_j\}$ (on s'arrête au premier non-rejet) ;

— les procédures **ascendantes** (*step-up*) : on part de la plus grande valeur p et on « remonte » ; on rejette $H_{(1)},\ldots,H_{(\hat{k})}$ où $\hat{k} = \max\{j : p_{(j)} \le \alpha_j\}$ (on s'arrête au premier rejet).

À seuils α_j égaux, la procédure ascendante rejette toujours au moins autant que la descendante, puisque $\hat{k} \ge k^* - 1$. Elle demande en contrepartie des conditions plus fortes pour garantir le contrôle.

Une procédure est dite **marginale** si elle n'utilise que les valeurs p brutes (Bonferroni, Holm, BH, …) et **conjointe** si elle exploite la loi conjointe des statistiques de test (minP, maxT). Les procédures conjointes tiennent compte de la dépendance et sont plus puissantes quand les tests sont fortement corrélés, au prix d'un calcul par ré-échantillonnage.

5 Contrôle du FWER

5.1 Bonferroni et Šidák

Proposition 5.1 (Bonferroni)**.** *La procédure qui rejette $H_0(j)$ si $p_j \le \alpha/m$ contrôle fortement le FWER au niveau $m_0\alpha/m \le \alpha$, sous n'importe quelle structure de dépendance. Sa valeur p ajustée est $\tilde{p}_j = \min(m\,p_j, 1)$.*

Démonstration. Par l'inégalité de Boole et la validité des valeurs p nulles, $\mathrm{FWER} = \mathbb{P}\big(\bigcup_{j\in\mathcal{H}_0}\{p_j \le \alpha/m\}\big) \le \sum_{j\in\mathcal{H}_0}\mathbb{P}(p_j \le \alpha/m) \le m_0\,\alpha/m$. Aucune hypothèse sur la dépendance n'intervient ; la borne vaut pour tout $\mathcal{H}_0$, d'où le contrôle fort. □

La procédure de Bonferroni est conservatrice pour deux raisons : la borne de Boole est lâche quand les événements de rejet ne sont pas disjoints (en particulier quand les tests sont positivement corrélés), et le niveau effectif est $m_0\alpha/m$ plutôt que α.

Si les statistiques sont indépendantes, on peut calculer exactement la loi de la plus petite valeur p. C'est l'idée de l'approche « minP » : on cherche un seuil commun c tel que $\mathbb{P}\{\min_j P_j \le c \mid H_0^C\} = \alpha$.

Proposition 5.2 (Šidák)**.** *Si les valeurs p des hypothèses nulles vraies sont indépendantes et uniformes, la procédure qui rejette $H_0(j)$ si $p_j \le c_m = 1-(1-\alpha)^{1/m}$ contrôle fortement le FWER au niveau $1-(1-\alpha)^{m_0/m} \le \alpha$, avec égalité sous H_0^C. Sa valeur p ajustée est $\tilde{p}_j = 1-(1-p_j)^m$.*

Démonstration. Par indépendance, $\mathbb{P}(V=0) = \mathbb{P}\big(\bigcap_{j\in\mathcal{H}_0}\{p_j > c\}\big) = \prod_{j\in\mathcal{H}_0}(1-c) = (1-c)^{m_0}$, donc $\mathrm{FWER} = 1-(1-c)^{m_0}$. Sous H_0^C, on résout $1-(1-c)^m = \alpha$, ce qui donne $c_m = 1-(1-\alpha)^{1/m}$. Pour $m_0 < m$, $\mathrm{FWER} = 1-(1-\alpha)^{m_0/m} \le \alpha$. Enfin, $p_j \le c_m \iff 1-(1-p_j)^m \le \alpha$, d'où la valeur p ajustée. □

Le gain par rapport à Bonferroni est minime : comme $(1-\alpha)^{1/m} \le 1-\alpha/m$ (inégalité de Bernoulli), on a $c_m \ge \alpha/m$, mais pour $\alpha = 0{,}05$ et $m = 1000$, $c_m = 5{,}13\times10^{-5}$ contre $5{,}00\times10^{-5}$. Šidák (1967)

a aussi montré que l'inégalité $\mathbb{P}(V = 0) \geq (1 - c)^{m_0}$ reste vraie pour des tests bilatéraux fondés sur un vecteur normal multivarié de corrélations quelconques, ce qui étend la validité de la procédure à ce cas.

5.2 Procédure descendante de Holm

On améliore Bonferroni sans aucune hypothèse supplémentaire en utilisant des seuils $\alpha_j = \alpha/(m - j + 1)$ dans une procédure descendante : $p_{(1)}$ est comparée à α/m, $p_{(2)}$ à $\alpha/(m-1)$, …, $p_{(m)}$ à α. Intuitivement, une fois $H_{(1)}$ rejetée, il ne reste au pire que $m - 1$ hypothèses nulles vraies possibles, et on peut corriger pour $m - 1$ tests.

Théorème 5.3 (Holm, 1979)**.** *La procédure descendante de Holm contrôle fortement le FWER au niveau α, sous n'importe quelle structure de dépendance. Ses valeurs p ajustées sont*

$$\tilde{p}_{(j)} = \max_{k \leq j} \ \min\bigl\{(m - k + 1)\, p_{(k)}, 1\bigr\}.$$

Démonstration. Fixons une configuration $\mathcal{H}_0$ avec $m_0 \geq 1$ (si $m_0 = 0$, il n'y a pas d'erreur possible) et considérons l'événement

$$A = \Bigl\{\min_{j \in \mathcal{H}_0} p_j > \alpha/m_0\Bigr\}.$$

Par l'inégalité de Boole, $\mathbb{P}(A^c) \leq \sum_{j \in \mathcal{H}_0} \mathbb{P}(p_j \leq \alpha/m_0) \leq \alpha$. Montrons que, sur A, la procédure ne rejette aucune hypothèse nulle vraie. Soit j^* le rang (dans l'ordre croissant des valeurs p) de la première hypothèse nulle vraie de la liste. Les $j^* - 1$ hypothèses qui la précèdent sont toutes fausses, donc $j^* - 1 \leq m_1 = m - m_0$, c'est-à-dire $m - j^* + 1 \geq m_0$. Le seuil à l'étape j^* vérifie donc $\alpha/(m - j^* + 1) \leq \alpha/m_0 < p_{(j^*)}$ sur A. La procédure s'arrête au plus tard à l'étape j^*, sans rejeter $H_{(j^*)}$ ni aucune hypothèse qui la suit. Ainsi $\{V > 0\} \subseteq A^c$ et FWER $\leq \alpha$.

Pour la valeur p ajustée : $H_{(j)}$ est rejetée au niveau α si et seulement si $p_{(k)} \leq \alpha/(m - k + 1)$ pour tout $k \leq j$, c'est-à-dire si $\max_{k \leq j}(m - k + 1)p_{(k)} \leq \alpha$. Le maximum garantit la monotonie $\tilde{p}_{(1)} \leq \cdots \leq \tilde{p}_{(m)}$. □

La procédure de Holm rejette toujours au moins autant d'hypothèses que celle de Bonferroni, puisque ses seuils sont plus grands, avec la même garantie. Il n'y a donc aucune raison d'utiliser Bonferroni plutôt que Holm, sinon la simplicité. La version descendante de Šidák utilise de même les seuils $1 - (1 - \alpha)^{1/(m-j+1)}$; sa valeur p ajustée est $\tilde{p}_{(j)} = \max_{k \leq j}\{1 - (1 - p_{(k)})^{m-k+1}\}$, et elle demande les mêmes conditions de dépendance que la procédure de Šidák à une étape.

Niveau doctorat : le principe de fermeture

La démonstration du théorème 5.3 est un cas particulier d'un principe général (Marcus, Peritz et Gabriel, 1976). Pour tout sous-ensemble $I \subseteq \{1, \ldots, m\}$, notons $H_I = \bigcap_{j \in I} H_0(j)$ l'hypothèse d'intersection, et supposons qu'on dispose pour chaque H_I d'un test « local » de niveau α. La *procédure fermée* rejette $H_0(j)$ si et seulement si toutes les intersections H_I avec $I \ni j$ sont rejetées par leur test local.

Elle contrôle fortement le FWER. En effet, soit $I_0 = \mathcal{H}_0$; pour commettre une erreur sur $j \in I_0$, il faut rejeter H_{I_0}, qui est vraie : cet événement a une probabilité au plus α.

Si le test local de H_I est le test de Bonferroni (rejeter si $\min_{j \in I} p_j \leq \alpha/|I|$), la procédure fermée se ramène exactement à la procédure de Holm (exercice). Avec le test de Simes comme

test local, on obtient la procédure de Hommel (1988), dont celle de Hochberg est un raccourci conservateur. Le même principe fonde les procédures hiérarchiques (*gatekeeping*) des essais cliniques.

5.3 Procédure ascendante de Hochberg

La procédure de Hochberg (1988) utilise les mêmes seuils $\alpha/(m-j+1)$ que Holm, mais de façon ascendante : on part de $p_{(m)}$ (comparée à α), puis $p_{(m-1)}$ (comparée à $\alpha/2$), et ainsi de suite ; au premier j tel que $p_{(j)} \le \alpha/(m-j+1)$, on rejette $H_{(1)},\dots,H_{(j)}$. Ses valeurs p ajustées sont

$$\tilde{p}_{(j)} = \min_{k \ge j} \ \min\bigl\{(m-k+1)\,p_{(k)}, 1\bigr\}.$$

Elle rejette toujours au moins autant que Holm. Sa validité repose sur l'*inégalité de Simes* (1986) : sous H_0^C,

$$\mathbb{P}\Bigl(\exists\, j : p_{(j)} \le \tfrac{j\alpha}{m}\Bigr) \le \alpha,$$

avec égalité quand les valeurs p sont indépendantes et uniformes. Cette inégalité est vraie sous indépendance et, plus généralement, sous une forme de dépendance positive (la propriété PRDS définie à la section 6 ; Sarkar, 1998), par exemple pour des tests unilatéraux fondés sur un vecteur normal multivarié à corrélations non négatives. Elle peut être fausse sous dépendance négative.

Conditions de la procédure de Hochberg

Contrairement à Holm, Hochberg *ne contrôle pas le FWER sous dépendance arbitraire* : on la réserve aux statistiques indépendantes ou PRDS. Le gain sur Holm est souvent faible ; Holm reste le choix par défaut quand la dépendance est inconnue.

5.4 Exemple numérique « papier »

Exemple 5.4. Une étude pilote compare l'expression de $m = 10$ gènes candidats entre deux groupes de patients. Les valeurs p ordonnées sont :

j	1	2	3	4	5	6	7	8	9	10
$p_{(j)}$	0,001	0,004	0,006	0,0072	0,008	0,028	0,041	0,120	0,380	0,750
$\alpha/(m-j+1)$	0,0050	0,0056	0,0063	0,0071	0,0083	0,0100	0,0125	0,0167	0,0250	0,0500
$j\alpha/m$	0,005	0,010	0,015	0,020	0,025	0,030	0,035	0,040	0,045	0,050

Avec $\alpha = 0{,}05$:

— *Bonferroni* : seuil 0,005 ; on rejette $H_{(1)}$ et $H_{(2)}$ (2 rejets). *Šidák* : seuil $1-0{,}95^{1/10} = 0{,}00512$; même résultat.

— *Holm* (descendante) : $0{,}001 \le 0{,}0050$, $0{,}004 \le 0{,}0056$, $0{,}006 \le 0{,}0063$, puis $0{,}0072 > 0{,}0071$: on s'arrête. On rejette $H_{(1)}$,$H_{(2)}$,$H_{(3)}$ (3 rejets), même si $p_{(5)} = 0{,}008$ est sous son seuil 0,0083 : une procédure descendante ne regarde plus rien après le premier non-rejet.

— *Hochberg* (ascendante) : on remonte depuis $j = 10$; $p_{(j)}$ dépasse son seuil pour $j = 10, \ldots, 6$; à $j = 5$, $0{,}008 \leq 0{,}0083$. On rejette $H_{(1)}, \ldots, H_{(5)}$ (5 rejets), y compris $H_{(4)}$, dont la valeur p dépasse pourtant son propre seuil.

Les valeurs p ajustées se calculent directement. Pour Holm, $\tilde{p}_{(4)} = \max\{10 \times 0{,}001,\ 9 \times 0{,}004,\ 8 \times 0{,}006,\ 7 \times 0{,}0072\} = 0{,}0504$; pour Hochberg, $\tilde{p}_{(4)} = \min\{7 \times 0{,}0072,\ 6 \times 0{,}008,\ 5 \times 0{,}028, \ldots\} = 0{,}048$. On reprendra cet exemple pour le FDR à la section 6.

5.5 Procédures conjointes minP et maxT

Quand les tests sont fortement corrélés, le nombre « effectif » de tests est bien inférieur à m, et les procédures marginales sont inutilement sévères. Les procédures conjointes utilisent plutôt la loi nulle d'un résumé de l'ensemble des statistiques.

Quantile commun ou seuil commun. Deux philosophies existent (Dudoit et van der Laan, p. 16).

— **maxT** (seuil commun) : on compare toutes les statistiques $|T_n(j)|$ au même seuil, le quantile $1 - \alpha$ de $\max_j |T_n(j)|$ sous l'hypothèse nulle. C'est naturel quand les statistiques ont la même loi marginale (p. ex. toutes $\mathcal{N}(0,1)$).

— **minP** (quantile commun) : on transforme d'abord chaque statistique en valeur p, puis on compare les valeurs p au quantile α de $\min_j P_j$. Chaque hypothèse est ainsi traitée « sur le même pied » même quand les lois marginales des statistiques diffèrent (degrés de liberté différents, valeurs manquantes, tests de natures différentes).

La procédure de Šidák est la procédure minP à une étape dans le cas particulier où les tests sont indépendants, puisqu'alors $\mathbb{P}(\min_j P_j \leq c) = 1 - (1 - c)^m$ est connue analytiquement.

Valeurs p ajustées. Pour la procédure maxT à une étape : $\tilde{p}_j = \mathbb{P}(\max_{1 \leq l \leq m} |T_l| \geq |t_j| \mid H_0^C)$, où t_j est la valeur observée et $(T_1, \ldots, T_m)$ suit la loi nulle conjointe. Pour la version *descendante* de Westfall et Young (1993), on ordonne les statistiques observées $|t_{r_1}| \geq |t_{r_2}| \geq \cdots \geq |t_{r_m}|$ et on pose

$$\tilde{p}_{r_j} = \max_{k \leq j} \mathbb{P}\Big(\max_{l \in \{r_k, \ldots, r_m\}} |T_l| \geq |t_{r_k}| \;\Big|\; H_0^C \Big).$$

Chaque hypothèse est comparée au maximum des statistiques des hypothèses *qui ne sont pas encore rejetées*, comme dans Holm. La version minP descendante s'écrit de la même façon avec $\min P_l$ et p_{r_k}. Si les tests sont indépendants, la procédure minP descendante est la procédure descendante de Šidák.

Algorithme de Westfall et Young (maxT descendante).

1. Calculer les statistiques observées et les ordonner : $|t_{r_1}| \geq \cdots \geq |t_{r_m}|$.
2. Pour $b = 1, \ldots, B$, générer un jeu de données sous l'hypothèse nulle (permutation des étiquettes de groupe, par exemple) et calculer $t_1^{(b)}, \ldots, t_m^{(b)}$.
3. Calculer les maxima successifs « par le bas » : $u_m^{(b)} = |t_{r_m}^{(b)}|$ et $u_j^{(b)} = \max(u_{j+1}^{(b)}, |t_{r_j}^{(b)}|)$.

4. Poser $\tilde{p}^*_{r_j} = B^{-1}\#\{b : u_j^{(b)} \geq |t_{r_j}|\}$, puis imposer la monotonie : $\tilde{p}_{r_j} = \max_{k \leq j} \tilde{p}^*_{r_k}$.

La version minP demande en plus une valeur p pour chaque statistique de chaque jeu ré-échantillonné, ce qui exige en principe un second niveau de ré-échantillonnage ; la fonction `mt.minP` utilise l'algorithme efficace de Ge, Dudoit et Speed (2003), qui évite cette double boucle.

Pivotalité des sous-ensembles. Les probabilités ci-dessus sont calculées sous H_0^C. Le contrôle est *fort* si la loi conjointe de $(T_j)_{j \in \mathcal{H}_0}$ ne dépend pas de la vérité des autres hypothèses (*pivotalité des sous-ensembles*, Westfall et Young) : c'est raisonnable pour des statistiques t de comparaison de moyennes, moins pour des tests de corrélations. L'approche de Dudoit et van der Laan (encadré plus bas) contourne cette condition.

5.6 Estimer la loi nulle : permutation ou bootstrap

La loi nulle conjointe est rarement connue ; on l'estime par ré-échantillonnage (Farcomeni, § 3.2.3).

Permutation. Pour comparer deux groupes, on permute les étiquettes de groupe en gardant intacts les vecteurs d'expression de chaque sujet, ce qui préserve la dépendance entre les gènes. Le test est *exact* en échantillon fini si les observations sont *échangeables* sous l'hypothèse nulle, c'est-à-dire si la loi conjointe des données est invariante par permutation des étiquettes. Pour la comparaison de deux groupes, cela veut dire que la loi *complète* (multivariée) est la même dans les deux groupes, pas seulement les moyennes. Si l'hypothèse nulle d'intérêt ne porte que sur les moyennes, l'échangeabilité peut échouer de deux façons (Dudoit et van der Laan, § 2.9) :

— *variances (ou moments supérieurs) inégales* : sous $H_0 : \mu_1 = \mu_2$, les données restent non échangeables si $\sigma_1 \neq \sigma_2$;

— *tailles inégales* : avec $n_1 = n_2$, la statistique de différence de moyennes a une loi de permutation qui reste approximativement correcte même si $\sigma_1 \neq \sigma_2$; avec $n_1 \neq n_2$ et des variances inégales, la loi de permutation est mal calibrée.

Pour la loi nulle conjointe, il faut de plus que la structure de corrélation entre gènes soit la même dans les deux groupes. Le code suivant illustre le problème pour un seul test, avec 2000 jeux de données simulés sous $H_0 : \mu_1 = \mu_2$ et $B = 999$ permutations.

```
## taux(n1, n2, s1, s2) : simule nsim = 2000 jeux de données sous H0, calcule pour
    chacun la valeur p
## de permutation (B = 999) de la différence de moyennes et de la statistique de Welch,
     et renvoie
## les taux de rejet à 5 % (code complet dans 11_tests_multiples_code.R)
set.seed(2027)
rbind("n1=n2=15, s1=s2=1" = taux(15, 15, 1, 1), "n1=n2=15, s1=4, s2=1" = taux(15, 15, 4,
     1),
      "n1=5, n2=25, s1=s2=1" = taux(5, 25, 1, 1), "n1=5, n2=25, s1=4, s2=1" = taux(5,
    25, 4, 1),
      "n1=5, n2=25, s1=1, s2=4" = taux(5, 25, 1, 4))
```

```
                          dif  welch
n1=n2=15, s1=s2=1      0.0510 0.0510
```

```
n1=n2=15, s1=4, s2=1    0.0505 0.0505
n1=5, n2=25, s1=s2=1    0.0510 0.0535
n1=5, n2=25, s1=4, s2=1 0.3465 0.0770
n1=5, n2=25, s1=1, s2=4 0.0000 0.0195
```

Avec des groupes de même taille, le niveau est respecté même si les variances diffèrent (les deux statistiques sont alors des fonctions monotones l'une de l'autre et donnent des tests identiques). Avec $n_1 = 5$, $n_2 = 25$ et la grande variance dans le petit groupe, le test de permutation de la différence de moyennes rejette 35 % du temps au lieu de 5 % ; dans la configuration inverse, il ne rejette jamais (perte de puissance). La statistique studentisée de Welch atténue beaucoup le problème, sans l'éliminer en petit échantillon ; on sait qu'elle donne un test de permutation asymptotiquement valide (Janssen, 1997). *Moralité* : utilisez une statistique studentisée et vérifiez l'équilibre des groupes avant de vous fier à une loi de permutation.

Bootstrap. Le bootstrap ré-échantillonne les sujets avec remise, dans chaque groupe, puis recentre les statistiques pour imposer l'hypothèse nulle. Il n'exige pas l'échangeabilité et s'applique plus largement (variances inégales, régression, corrélations), mais ne garantit qu'un contrôle *asymptotique*, alors que la permutation donne un contrôle exact quand l'échangeabilité tient (Dudoit et van der Laan, procédure 3.6, p. 118, et procédure 4.1, p. 184–185).

Niveau doctorat : la loi nulle de Dudoit et van der Laan

Dudoit et van der Laan (2008, ch. 2) proposent de ne pas spécifier une loi des *données* sous l'hypothèse nulle (ce qui exige la pivotalité des sous-ensembles), mais directement une loi nulle Q_0 des *statistiques de test* qui domine asymptotiquement la loi des statistiques des hypothèses nulles vraies : pour tout x,

$$\limsup_{n\to\infty} \mathbb{P}\Big(\max_{j\in\mathcal{H}_0} T_n(j) > x\Big) \leq \mathbb{P}_{Q_0}\Big(\max_{j\in\mathcal{H}_0} Z(j) > x\Big).$$

Leur choix par défaut est la loi limite du vecteur « translaté et réduit »

$$Z_n(j) = \sqrt{\min\Big(1, \tfrac{\tau_0(j)}{\operatorname{Var} T_n(j)}\Big)}\,\{T_n(j) - \mathrm{E}[T_n(j)]\} + \lambda_0(j),$$

où $\lambda_0(j)$ et $\tau_0(j)$ sont des bornes pour la moyenne et la variance de $T_n(j)$ sous $H_0(j)$ (pour des statistiques t, $\lambda_0 = 0$ et $\tau_0 = 1$). La loi de Z_n est estimée par bootstrap non paramétrique, en remplaçant $\mathrm{E}[T_n(j)]$ et $\operatorname{Var} T_n(j)$ par leurs analogues bootstrap : c'est l'option `nulldist = "boot.cs"` (« centrée et réduite ») de la fonction `MTP`. Les statistiques des hypothèses fausses n'ont pas besoin d'avoir la bonne loi : le recentrage les ramène vers la loi nulle, et la procédure (maxT, minP, à une étape ou descendante) appliquée avec Q_0 contrôle asymptotiquement le FWER au sens fort, sans condition de pivotalité. Le prix à payer est le caractère seulement asymptotique de ce contrôle ; avec $n = 10$ sujets par groupe, la loi bootstrap des statistiques t a souvent des queues trop lourdes, ce qui rend la procédure conservatrice (voir l'exemple 6.5). Les mêmes auteurs construisent des procédures pour le gFWER et le TPPFP par *augmentation* : on part des R_0 rejets d'une procédure qui contrôle le FWER et on ajoute les k hypothèses suivantes (gFWER(k)) ou les $\lfloor qR_0/(1-q)\rfloor$ suivantes (TPPFP(q)) dans l'ordre des valeurs p (van der Laan, Dudoit et Pollard, 2004 ; fonctions `fwer2gfwer` et `fwer2tppfp`).

À retenir

— Bonferroni et Holm contrôlent fortement le FWER sous *toute* dépendance ; Holm domine Bonferroni.

— Šidák et Hochberg demandent l'indépendance ou une dépendance positive.

— minP/maxT exploitent la dépendance par la loi nulle conjointe ; version descendante de Westfall–Young plus puissante que la version à une étape.

— Permutation : contrôle exact, mais exige l'échangeabilité (attention aux tailles *et* variances inégales). Bootstrap : plus général, contrôle asymptotique seulement.

6 Contrôle du FDR

Le FDR, introduit par Benjamini et Hochberg (1995), est devenu le critère le plus utilisé pour les données de haut débit (Farcomeni, § 3.5.2).

6.1 La procédure de Benjamini et Hochberg

Définition 6.1 (Procédure BH)**.** Soit $\hat{k} = \max\{j : p_{(j)} \leq j\alpha/m\}$ (avec $\hat{k} = 0$ si l'ensemble est vide). On rejette $H_{(1)}, \ldots, H_{(\hat{k})}$. C'est une procédure ascendante avec seuils $\alpha_j = j\alpha/m$. Ses valeurs p ajustées sont

$$\tilde{p}_{(j)} = \min_{k \geq j} \ \min\Big\{\frac{m\, p_{(k)}}{k}, 1\Big\}.$$

Le premier seuil, α/m, est celui de Bonferroni ; le dernier, α, est celui d'un test individuel. Les seuils de BH croissent linéairement en j, alors que ceux de Holm ou Hochberg restent de l'ordre de α/m pour les petits rangs (figure 2). Plus le seuil est élevé, plus la procédure est puissante.

Une interprétation utile : si on rejette toutes les hypothèses avec $p_j \leq t$, le nombre attendu de faux rejets est au plus $m_0 t \leq mt$, et le nombre de rejets est $R(t) = \#\{j : p_j \leq t\}$. L'estimateur naturel de la FDP est donc $\widehat{\mathrm{FDR}}(t) = mt/R(t)$, et la procédure BH revient à prendre le plus grand seuil t de la forme $p_{(j)}$ tel que $\widehat{\mathrm{FDR}}(t) \leq \alpha$.

Théorème 6.2 (Benjamini et Hochberg, 1995)**.** *Si les valeurs p des hypothèses nulles vraies sont indépendantes entre elles et indépendantes de celles des hypothèses fausses, et si elles suivent la loi uniforme, alors la procédure BH vérifie*

$$\mathrm{FDR} = \frac{m_0}{m}\,\alpha = \pi_0\,\alpha \leq \alpha.$$

Si les valeurs p nulles sont seulement valides ($\mathbb{P}(p_j \leq u) \leq u$), on a $\mathrm{FDR} \leq \pi_0\alpha$.

Deux conséquences importantes. D'abord, le contrôle est fort (il vaut pour tout $\mathcal{H}_0$). Ensuite, BH contrôle le FDR au niveau $\pi_0\alpha$ et non α : la procédure est conservatrice d'un facteur π_0, parce qu'elle fait comme si toutes les hypothèses étaient vraies (elle remplace m_0 par m dans $\widehat{\mathrm{FDR}}$). C'est ce qui motive l'estimation de π_0 (§ 6.3).

Niveau doctorat : démonstration du théorème 6.2

Pour $i \in \{1,\dots,m\}$, notons $R^{(i)}$ le nombre de rejets de la procédure BH appliquée au vecteur de valeurs p où p_i est remplacée par 0. C'est une fonction de $p_{-i} = (p_l)_{l\neq i}$ seulement, et $R^{(i)} \geq 1$.
Lemme. $H_0(i)$ est rejetée par BH si et seulement si $p_i \leq \alpha R^{(i)}/m$, et dans ce cas $R = R^{(i)}$.
Preuve du lemme. Diminuer p_i ne peut qu'augmenter le nombre de valeurs p sous chaque seuil, donc $\hat{k} \leq R^{(i)}$. Si $p_i \leq \alpha R^{(i)}/m$: dans le vecteur modifié, les $R^{(i)}$ plus petites valeurs p (dont 0) sont $\leq \alpha R^{(i)}/m$; dans le vecteur original, p_i l'est aussi, de sorte qu'au moins $R^{(i)}$ valeurs p sont $\leq \alpha R^{(i)}/m$, donc $p_{(R^{(i)})} \leq \alpha R^{(i)}/m$ et $\hat{k} \geq R^{(i)}$. Ainsi $\hat{k} = R^{(i)}$ et $p_i \leq \alpha\hat{k}/m$: $H_0(i)$ est rejetée. Si $p_i > \alpha R^{(i)}/m \geq \alpha\hat{k}/m$, elle ne l'est pas. □
Preuve du théorème. Par le lemme,

$$\mathrm{FDP} = \sum_{i\in\mathcal{H}_0} \frac{\mathbf{1}\{H_0(i) \text{ rejetée}\}}{R \vee 1} = \sum_{i\in\mathcal{H}_0} \frac{\mathbf{1}\{p_i \leq \alpha R^{(i)}/m\}}{R^{(i)}}.$$

Pour $i \in \mathcal{H}_0$, p_i est indépendante de p_{-i}, donc de $R^{(i)}$. En conditionnant sur p_{-i} :

$$\mathrm{E}\Big[\frac{\mathbf{1}\{p_i \leq \alpha R^{(i)}/m\}}{R^{(i)}} \,\Big|\, p_{-i}\Big] = \frac{\mathbb{P}(p_i \leq \alpha R^{(i)}/m \mid p_{-i})}{R^{(i)}} = \frac{\alpha R^{(i)}/m}{R^{(i)}} = \frac{\alpha}{m},$$

avec « $\leq$ » si p_i est seulement valide. En sommant sur les m_0 hypothèses nulles vraies, $\mathrm{FDR} = m_0\alpha/m$. Remarquez qu'on a utilisé l'indépendance entre p_i ($i \in \mathcal{H}_0$) et *toutes* les autres valeurs p.
Dépendance arbitraire (BY). Sans indépendance, on a quand même, pour la procédure ascendante de seuils $j\alpha'/m$: si $H_0(i)$ est rejetée et $p_i \in \big((k-1)\alpha'/m, k\alpha'/m\big]$, alors $R \geq k$ puisque $p_i \leq \alpha' R/m$. Donc

$$\frac{\mathbf{1}\{H_0(i) \text{ rejetée}\}}{R \vee 1} \leq \sum_{k=1}^{m} \frac{1}{k}\,\mathbf{1}\Big\{\tfrac{(k-1)\alpha'}{m} < p_i \leq \tfrac{k\alpha'}{m}\Big\},$$

et l'espérance du membre de droite est au plus $\frac{\alpha'}{m}\sum_{k=1}^{m}\frac{1}{k}$. En sommant sur $\mathcal{H}_0$: $\mathrm{FDR} \leq \frac{m_0}{m}\,\alpha' c(m)$ avec $c(m) = \sum_{k=1}^{m} 1/k$. Le choix $\alpha' = \alpha/c(m)$ donne la procédure de Benjamini et Yekutieli (2001), qui contrôle le FDR à $\pi_0\alpha$ sous n'importe quelle dépendance.

6.2 Dépendance : PRDS et procédure de Benjamini–Yekutieli

L'indépendance est rarement plausible pour des gènes (voies métaboliques communes, effets de lot). Deux résultats permettent de s'en affranchir (Farcomeni, § 3.6.2).

Définition 6.3 (PRDS)**.** Un vecteur $X = (X_1,\dots,X_m)$ de statistiques (ou de valeurs p) est *positivement dépendant par régression sur le sous-ensemble* $\mathcal{H}_0$ (PRDS) si, pour tout ensemble croissant D (si $x \in D$ et $y \geq x$ composante par composante, alors $y \in D$) et tout $i \in \mathcal{H}_0$, la fonction $x \mapsto \mathbb{P}(X \in D \mid X_i = x)$ est non décroissante.

Benjamini et Yekutieli (2001) montrent que BH contrôle encore le FDR au niveau $\pi_0\alpha$ si les valeurs p sont PRDS sur $\mathcal{H}_0$. L'exemple type est celui de statistiques normales multivariées avec corrélations non négatives et des tests *unilatéraux*. Pour des tests bilatéraux à corrélations quelconques, aucun

résultat général n'est établi, mais les nombreuses études de simulation montrent que BH reste très proche du niveau nominal.

Sous dépendance *arbitraire*, la procédure de **Benjamini–Yekutieli** (BY) remplace α par $\alpha/c(m)$, avec $c(m) = \sum_{j=1}^{m} 1/j \approx \ln m + 0{,}577$; ses seuils sont $j\alpha/\{m\,c(m)\}$ (démonstration dans l'encadré précédent). Pour $m = 1000$, $c(m) \approx 7{,}49$: la perte de puissance est importante, et BY est souvent moins puissante que Holm quand peu d'hypothèses sont fausses (figure 2). On la réserve aux situations où une dépendance négative est plausible. Enfin, la procédure de ré-échantillonnage de Yekutieli et Benjamini (1999) permet une dépendance quelconque entre les hypothèses nulles vraies, mais exige que celles-ci soient indépendantes des hypothèses fausses (Dudoit et van der Laan, p. 148).

Exemple 6.4 (Suite de l'exemple 5.4)**.** Avec les seuils $j\alpha/m = 0{,}005\,j$: on remonte depuis $j = 10$; $0{,}750 > 0{,}050$, $0{,}380 > 0{,}045$, $0{,}120 > 0{,}040$, $0{,}041 > 0{,}035$, puis $p_{(6)} = 0{,}028 \leq 0{,}030$. La procédure BH rejette $H_{(1)}, \ldots, H_{(6)}$ (6 rejets, contre 3 pour Holm et 5 pour Hochberg). Pour BY, $c(10) = 2{,}929$ et les seuils sont $0{,}00171\,j$: $p_{(6)} = 0{,}028 > 0{,}0102$, mais $p_{(5)} = 0{,}008 \leq 0{,}0085$, d'où 5 rejets. La vérification avec `multtest` :

```
library(multtest)
p <- c(0.001, 0.004, 0.006, 0.0072, 0.008, 0.028, 0.041, 0.120, 0.380, 0.750)
res <- mt.rawp2adjp(p, proc = c("Bonferroni", "SidakSS", "Holm", "Hochberg", "BH", "BY")
    )
round(res$adjp, 4); colSums(res$adjp <= 0.05)
```

```
      rawp Bonferroni SidakSS   Holm Hochberg     BH     BY
 [1,] 0.0010      0.010  0.0100 0.0100    0.010 0.0100 0.0293
 [2,] 0.0040      0.040  0.0393 0.0360    0.036 0.0160 0.0469
 [3,] 0.0060      0.060  0.0584 0.0480    0.048 0.0160 0.0469
 [4,] 0.0072      0.072  0.0697 0.0504    0.048 0.0160 0.0469
 [5,] 0.0080      0.080  0.0772 0.0504    0.048 0.0160 0.0469
  ...
    rawp Bonferroni    SidakSS       Holm   Hochberg         BH         BY
       7          2          2          3          5          6          5
```

Attention : `mt.rawp2adjp` renvoie les valeurs p ajustées *dans l'ordre croissant des valeurs p brutes* ; l'élément `index` permet de revenir à l'ordre original (voir l'exemple suivant).

6.3 Estimer la proportion d'hypothèses nulles vraies

Puisque BH contrôle le FDR à $\pi_0\alpha$, on récupère de la puissance en appliquant BH au niveau $\alpha/\hat{\pi}_0$, c'est-à-dire avec les seuils $j\alpha/(m\hat{\pi}_0) = j\alpha/\hat{m}_0$ (procédure « adaptative » ou « plug-in »). Farcomeni écrit $a = 1-\pi_0$ pour la proportion d'hypothèses nulles *fausses* et les seuils $j\alpha/\{m(1-\hat{a})\}$ (Farcomeni, § 3.4.3 ; Dudoit et van der Laan, p. 297). Quelques estimateurs :

— **Storey (2002)** : pour un $\lambda \in (0,1)$ fixé (souvent 0,5),

$$\hat{\pi}_0(\lambda) = \frac{\#\{j : p_j > \lambda\}}{m(1-\lambda)}.$$

Les valeurs p des hypothèses vraies sont uniformes, celles des hypothèses fausses sont concentrées près de 0 ; au-delà de λ, on ne voit (presque) que des hypothèses vraies, et $\mathrm{E}\,\#\{p_j > \lambda\} \geq m_0(1-\lambda)$. L'estimateur est donc biaisé vers le haut, ce qui est le côté prudent. Le choix de λ arbitre entre biais (petit λ) et variance (grand λ).

— **Benjamini et Hochberg (2000)**, « pente la plus faible » (option `"ABH"` de `mt.rawp2adjp`) : on estime m_0 par la pente de la droite reliant le point $(1, m+1)$ aux points $(p_{(j)}, j)$ du graphique de j en fonction de $p_{(j)}$.

— **Benjamini, Krieger et Yekutieli (2006)**, procédure *à deux étapes* (option `"TSBH"`) : (1) appliquer BH au niveau $\alpha' = \alpha/(1+\alpha)$ et compter les rejets r_1 ; (2) poser $\hat{m}_0 = m - r_1$ et appliquer BH au niveau $\alpha' m/\hat{m}_0$. Cette procédure contrôle le FDR à α sous indépendance, en échantillon fini.

Perdre le contrôle du FDR

L'estimateur $\hat{\pi}_0$ est aléatoire. S'il sous-estime π_0, la procédure adaptative utilise des seuils trop grands, et le FDR peut dépasser α. Le risque est réel quand m est petit (quelques dizaines d'hypothèses), quand les échantillons sont petits (valeurs p mal calibrées) ou quand les tests sont fortement dépendants (la distribution empirique des valeurs p fluctue beaucoup). Parmi les remèdes proposés (Farcomeni, § 3.4.3) : estimer m_1 de façon prudente à partir d'une procédure qui contrôle le FWER (on compte comme fausses seulement les hypothèses rejetées par Holm, p. ex.), ou utiliser des bornes uniformes sur la fonction de répartition empirique des valeurs p pour garantir que $\hat{\pi}_0 \geq \pi_0$ avec grande probabilité. La procédure de Benjamini–Krieger–Yekutieli est conçue pour garder le contrôle exact sous indépendance.

6.4 Exemple en R : données simulées avec dépendance

Exemple 6.5. On simule $m = 1000$ gènes pour $n_1 = n_2 = 10$ sujets. Les gènes sont regroupés en blocs de 20 qui partagent un facteur latent (corrélation 0,5 à l'intérieur d'un bloc). Les 100 premiers gènes sont différentiellement exprimés (différence de moyennes de 2 pour 50 gènes, de 1,25 pour 50 autres ; écart type 1). On connaît donc la vérité, ce qui permet de compter V et S.

```
library(multtest)
set.seed(7360)
m <- 1000; n1 <- 10; n2 <- 10; taille.bloc <- 20; rho <- 0.5; m1 <- 100
delta <- c(rep(2, m1/2), rep(1.25, m1/2), rep(0, m - m1))
gen <- function(n) {   # facteur commun par bloc + bruit propre au gène
  f <- matrix(rnorm(n * m / taille.bloc), n)[, rep(1:(m / taille.bloc), each = taille.
    bloc)]
  t(sqrt(rho) * f + sqrt(1 - rho) * matrix(rnorm(n * m), n))
}
X <- cbind(gen(n1), gen(n2) + delta)            # matrice m x (n1 + n2)
cl <- rep(0:1, c(n1, n2)); h0 <- delta == 0
tw <- mt.teststat(X, cl, test = "t")            # statistique de Welch
rawp <- 2 * pt(-abs(tw), df = n1 + n2 - 2)
procs <- c("Bonferroni", "Holm", "Hochberg", "SidakSS", "SidakSD", "BH", "BY", "ABH", "
    TSBH")
res <- mt.rawp2adjp(rawp, proc = procs, alpha = 0.05)
adjp <- res$adjp[order(res$index), ]            # retour à l'ordre des gènes
rej <- adjp <= 0.05
rbind(R = colSums(rej), V = colSums(rej & h0), S = colSums(rej & !h0))
res$h0.ABH; res$h0.TSBH; sum(rawp > 0.5) / (m * 0.5)   # estimations de m0 et pi0
```

```
  rawp Bonferroni Holm Hochberg SidakSS SidakSD BH BY ABH TSBH_0.05
R  145         30   30       30      30      30 76 47  78        78
```

```
V   56          0    0        0       0       0  9  2  10        10
S   89         30   30       30      30      30 67 45  68        68
[1] 915
h0.TSBH_0.05
         924
[1] 0.894
```

Avec des groupes de même taille, la statistique de Welch coïncide avec la statistique t à variance commune, d'où la loi de Student à 18 degrés de liberté. Sans correction, on obtient 145 rejets dont 56 faux ($\approx m_0\alpha = 45$ en moyenne). Les procédures du FWER ne font aucune erreur mais ne trouvent que 30 des 100 gènes. BH en trouve 67, au prix de 9 faux positifs : la FDP observée vaut $9/76 = 11{,}8\,\%$. Ce n'est pas une contradiction : le FDR est une *espérance*, et la FDP d'une réalisation fluctue autour d'elle, d'autant plus que les tests sont dépendants (les faux positifs arrivent par blocs). Les procédures adaptatives estiment m_0 à 915 ou 924 (vraie valeur 900), Storey donne $\hat{\pi}_0 = 0{,}894$; le gain sur BH est faible puisque π_0 est proche de 1.

Le graphique du nombre de rejets en fonction du niveau nominal (figure 1) résume la comparaison pour tous les niveaux à la fois ; il est produit directement par `mt.plot`.

```
choix <- c("rawp", "Bonferroni", "Holm", "Hochberg", "BH", "BY", "TSBH_0.05")
mt.plot(adjp[, choix], tw, plottype = "rvsa", proc = choix,
        alpha = seq(0, 0.2, length = 401), leg = c(0.003, 262),
        col = c(1, 2, 2, 2, 4, 4, 3), lty = c(1, 1, 2, 3, 1, 3, 2), lwd = 2, ylim = c(0,
    260))
```

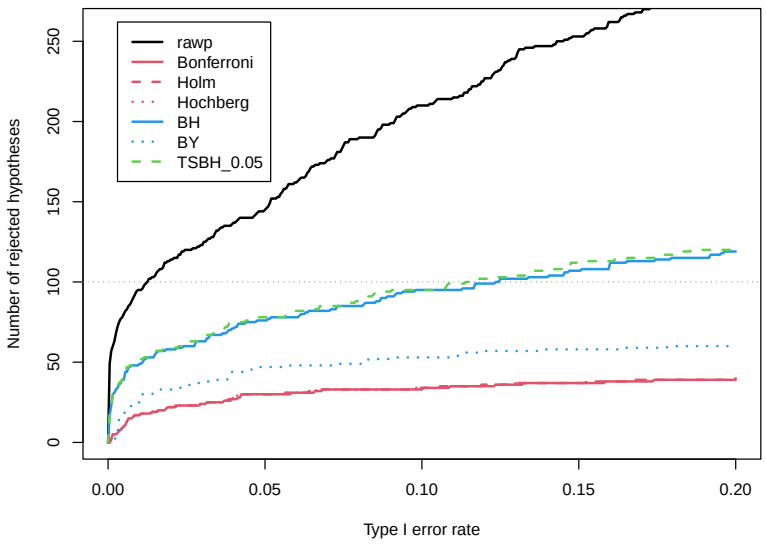


FIGURE 1 – Nombre d'hypothèses rejetées en fonction du niveau nominal α (FWER pour Bonferroni, Holm, Hochberg ; FDR pour BH, BY, TSBH), données simulées de l'exemple 6.5. La ligne pointillée grise indique les $m_1 = 100$ gènes réellement différentiellement exprimés ; la courbe `rawp` correspond aux tests sans correction. Les trois procédures du FWER sont confondues.

La figure 2 montre pourquoi BH rejette plus : pour les premiers rangs, les seuils sont semblables, mais le seuil $j\alpha/m$ croît avec le rang alors que $\alpha/(m-j+1)$ reste pratiquement égal à α/m tant que $j \ll m$. Le seuil de BY est environ 7,5 fois plus petit que celui de BH.

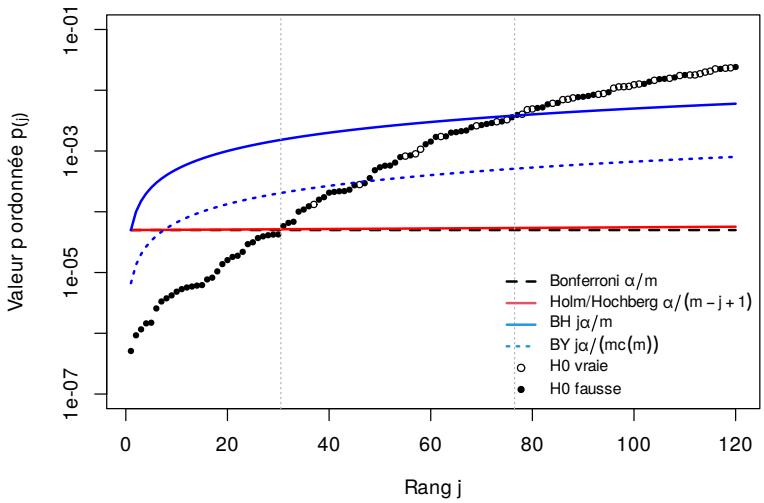


FIGURE 2 – Les 120 plus petites valeurs p ordonnées de l'exemple 6.5 (échelle logarithmique) et les seuils des procédures de Bonferroni, Holm/Hochberg, BH et BY pour $\alpha = 0{,}05$. Les lignes verticales indiquent les nombres de rejets de Holm (30) et de BH (76). Points pleins : gènes différentiellement exprimés.

Procédures conjointes sur les mêmes données. Les fonctions `mt.maxT` et `mt.minP` appliquent les procédures descendantes de Westfall–Young avec une loi nulle par permutation des étiquettes, et `MTP` offre le cadre général de Dudoit et van der Laan (ici bootstrap centré et réduit). Les colonnes `rawp` de `mt.maxT` sont des valeurs p de permutation, de résolution $1/B$.

```
maxT <- mt.maxT(X, cl, test = "t", side = "abs", B = 2000)
minP <- mt.minP(X, cl, test = "t", side = "abs", B = 2000)
head(maxT, 3); head(minP, 3)
minP50 <- mt.minP(X, cl, test = "t", side = "abs", B = 50000)
c(maxT = sum(maxT$adjp <= 0.05), minP = sum(minP$adjp <= 0.05),
  holm.perm = sum(p.adjust(maxT$rawp, "holm") <= 0.05), minP.50000 = sum(minP50$adjp <=
    0.05))
boot <- MTP(X = X, Y = cl, test = "t.twosamp.unequalvar", alternative = "two.sided",
            typeone = "fwer", alpha = 0.05, nulldist = "boot.cs", B = 1000,
            method = "sd.maxT", seed = 2027)
summary(boot)
```

```
   index teststat  rawp   adjp
21    21 7.593348 5e-04 0.0030
24    24 7.272760 1e-03 0.0035
50    50 7.153449 5e-04 0.0035
   index teststat  rawp   adjp plower
21    21 7.593348 5e-04 0.3445  5e-04
50    50 7.153449 5e-04 0.3445  5e-04
55    55 7.034096 5e-04 0.3445  5e-04
      maxT       minP  holm.perm minP.50000
        30          0          0         20
MTP:  sd.maxT
Type I error rate:  fwer
  Level Rejections
1  0.05         15
```

Trois leçons. (1) La procédure maxT de permutation rejette 30 hypothèses (aucune fausse), autant que Holm : avec une dépendance par blocs modérée et m grand, le gain conjoint est faible. (2) La procédure minP, et plus généralement toute procédure appliquée à des valeurs p de permutation, est limitée par la *granularité* : avec $B = 2000$, la plus petite valeur p possible est $1/2000$, beaucoup plus grande que $\alpha/m = 5 \times 10^{-5}$; il faut $B \gg m/\alpha$ ($B = 50\,000$ donne 20 rejets). La procédure maxT n'a pas ce problème, car elle compare les statistiques elles-mêmes au maximum ré-échantillonné. (3) Le bootstrap centré et réduit est plus conservateur (15 rejets) avec 10 sujets par groupe, comme annoncé dans l'encadré « Niveau doctorat » ; son contrôle n'est qu'asymptotique, mais il ne demande ni échangeabilité ni pivotalité.

6.5 Taux d'erreur empiriques : une petite simulation

Pour vérifier le théorème 6.2 et l'effet de la dépendance, on répète 2000 fois une expérience avec $m = 200$ gènes, dont 20 différentiellement exprimés ($\delta = 1{,}5$), $n = 10$ par groupe, avec des blocs de 20 gènes indépendants ($\rho = 0$) ou fortement corrélés ($\rho = 0{,}8$). On estime le FWER, le FDR (moyenne des FDP) et la puissance moyenne. La fonction `une.sim`, comme tout le code de la section, se trouve dans le fichier `11_tests_multiples_code.R`.

```
for (rho in c(0, 0.8)) {   # une.sim() : génère les données et applique p.adjust()
  out <- replicate(2000, une.sim(rho)); print(round(apply(out, 1:2, mean), 3))
}
```

```
rho = 0
      bonferroni  holm hochberg    BH    BY
FWER       0.043 0.044    0.044 0.370 0.033
FDP        0.011 0.011    0.011 0.045 0.008
puiss      0.184 0.185    0.185 0.460 0.151
rho = 0.8
      bonferroni  holm hochberg    BH    BY
FWER       0.024 0.025    0.025 0.158 0.020
FDP        0.015 0.015    0.015 0.032 0.005
puiss      0.194 0.198    0.198 0.472 0.216
```

Sous indépendance, le FDR de BH vaut $0{,}045 = \pi_0\alpha = 0{,}9 \times 0{,}05$, exactement comme le prédit le théorème 6.2 ; BH multiplie la puissance par 2,5 par rapport à Holm, mais son FWER atteint 37 % : BH ne contrôle pas le FWER. Avec une forte dépendance positive, toutes les procédures deviennent conservatrices pour leur propre critère (Bonferroni et Holm à 2,5 %, BH à 3,2 %), ce qui illustre le résultat PRDS et le coût de Bonferroni sous corrélation.

À retenir

- BH : procédure ascendante, seuils $j\alpha/m$; FDR $= \pi_0\alpha$ sous indépendance, $\leq \pi_0\alpha$ sous PRDS.
- BY : seuils $j\alpha/\{m\,c(m)\}$, $c(m) = \sum 1/j$; valide sous toute dépendance, mais peu puissante.
- Procédures adaptatives : seuils $j\alpha/\hat{m}_0$; plus puissantes quand π_0 est nettement inférieure à 1, mais risque de perdre le contrôle si $\hat{\pi}_0$ est instable.
- Le FDR contrôle la FDP *en moyenne* : la FDP d'une étude donnée peut dépasser α.

Compléments. Le contrôle du FDX(q) (Farcomeni, § 3.5.3) borne la probabilité que la FDP dépasse q plutôt que sa moyenne ; *ce sujet n'est pas au contenu du cours* (voir l'encadré « Niveau doctorat » de la section 5.6 et `typeone = "tppfp"` dans `MTP`). Sous une dépendance faible, les procédures conçues sous indépendance contrôlent asymptotiquement leur critère quand $m \to \infty$ (Farcomeni, § 3.6).

7 Choisir un critère, et ce qu'on peut en attendre

7.1 FWER ou FDR ?

Le choix dépend de ce qu'on fera des hypothèses rejetées.

— **Étude confirmatoire** (essai clinique avec quelques critères d'évaluation secondaires, sous-groupes pré-spécifiés, comparaisons multiples de doses) : chaque rejet sera interprété comme une conclusion et pourra mener à une décision. Un seul faux positif peut être coûteux : on contrôle le **FWER** (Holm par défaut ; Hochberg si dépendance positive ; procédure fermée si les hypothèses sont hiérarchisées). Le nombre de tests est généralement modéré.

— **Criblage suivi de validation** (micropuces, études d'association pangénomiques, criblage de molécules) : la liste des hypothèses rejetées sert à choisir les candidats qui seront vérifiés par une expérience indépendante, à plus petite échelle (PCR quantitative, réplication dans une autre cohorte). On tolère une proportion connue de faux positifs pour ne pas manquer de vrais signaux : on contrôle le **FDR** (BH, ou une version adaptative ; BY si la dépendance peut être négative) ou le FDX (Farcomeni, § 3.7).

Dans les deux cas, on fixe le critère et la procédure *avant* de voir les données, et on rapporte le nombre total d'hypothèses testées.

Niveau d'une liste et niveau d'une hypothèse

Le FDR est une propriété de la *liste* rejetée : une valeur q de 0,03 pour le gène j ne veut pas dire que ce gène a 3 % de chances d'être un faux positif ; les gènes du bas de la liste sont bien plus souvent nuls que ceux du haut (FDR local, hors contenu).

7.2 Exemple traité : nombres attendus de rejets

Exemple 7.1 (Examen final H26, question 3)**.** On teste $m = 1000$ hypothèses nulles d'égalité de moyennes entre deux groupes ; $m_0 = 980$ sont vraies et $m_1 = 20$ fausses. Pour une hypothèse fausse, la puissance du test est 0,9 au seuil 0,05, 0,75 au seuil qui contrôle le FDR à 0,05 et 0,6 au seuil 0,00005. On cherche $\mathrm{E}[R]$ et $\mathrm{E}[V] = \mathrm{E}[N_{1|0}]$ pour trois stratégies.

(a) Chaque hypothèse au seuil 0,05. Par linéarité de l'espérance, $\mathrm{E}[V] = 0{,}05 \times 980 = 49$ et $\mathrm{E}[S] = 0{,}9 \times 20 = 18$, donc $\mathrm{E}[R] = 49 + 18 = 67$. Environ 73 % des rejets seraient des faux positifs.

(b) Contrôle du FDR à 0,05. BH rejette toutes les hypothèses sous un seuil aléatoire t. En approximant le FDR par $\mathrm{E}[V]/\mathrm{E}[R]$ (justifié quand m est grand, car V/m et R/m se concentrent autour de leurs

espérances) et en notant t le seuil effectif :

$$0{,}05 = \frac{980\,t}{980\,t + 0{,}75 \times 20} \iff 0{,}95 \times 980\,t = 0{,}05 \times 15 \iff t = \frac{0{,}75}{931} = 0{,}000806.$$

Donc $\mathrm{E}[V] = 980\,t = 0{,}79$ et $\mathrm{E}[R] = 0{,}79 + 15 = 15{,}8$ (on vérifie : $0{,}79/15{,}8 = 0{,}05$).

(c) Contrôle du FWER à 0,05. Le seuil $0{,}00005 = 0{,}05/1000$ est celui de Bonferroni. Alors $\mathrm{E}[V] = 980 \times 0{,}00005 = 0{,}049$; c'est aussi une borne supérieure du FWER (proposition 3.3), qui est donc d'environ 0,05. Enfin $\mathrm{E}[R] = 0{,}049 + 0{,}6 \times 20 = 12{,}05$.

Bilan. De (a) à (b), on élimine 98 % des faux positifs attendus en perdant 3 vraies découvertes sur 18 ; de (b) à (c), on évite moins d'un faux positif pour 3 vraies découvertes de moins.

8 Application : découverte de gènes par micropuces

Farcomeni (§ 3.7) applique les procédures à deux jeux de données classiques.

Cancer du côlon (Alon et coll., 1999). Environ 2000 gènes mesurés sur 62 tissus (40 tumeurs, 22 tissus normaux). L'histogramme des valeurs p brutes montre un excès marqué près de 0 par rapport à la loi uniforme attendue si toutes les hypothèses étaient vraies : c'est le premier diagnostic à faire, car il indique que $\pi_0 < 1$ et donne une idée de $\hat{\pi}_0$ (la hauteur de la partie plate de l'histogramme). Les procédures du FDR, surtout adaptatives, détectent nettement plus de gènes que celles du FWER.

Leucémies ALL et AML (Golub et coll., 1999). Le jeu de données `golub` de `multtest` contient 3051 gènes pour 38 patients (27 ALL, 11 AML). Farcomeni utilise les procédures d'inférence simultanée pour *sélectionner* les gènes qui serviront de variables explicatives dans une règle de classification (par exemple les k plus proches voisins). C'est raisonnable quand l'échantillon est petit et que quelques gènes très différenciés suffisent à séparer les groupes. (L'analyse complète de ces données avec les procédures du tableau 3.5 de Farcomeni fait l'objet du devoir 2 ; elle n'est pas reprise ici.)

Tests multiples et sélection de variables pour la classification

Sélectionner des variables par tests multiples n'est *pas* une stratégie recommandée en général. Le test répond à la question « la moyenne diffère-t-elle ? », pas à « la variable aide-t-elle à prédire ? ». Quand l'échantillon est grand, des gènes dont l'effet individuel est trop faible pour être déclaré significatif après correction peuvent quand même contribuer à la classification, surtout combinés ; inversement, des gènes très significatifs peuvent être redondants. De plus, si la sélection est faite sur toutes les données avant d'estimer l'erreur de classification, cette erreur est optimiste : la sélection doit être refaite à l'intérieur de chaque pli de la validation croisée. On préfère des méthodes de prédiction qui sélectionnent ou pénalisent les variables selon leur apport prédictif (lasso, forêts aléatoires, etc.).

À retenir : démarche pour une étude de haut débit

1. Choisir le critère selon l'usage des résultats : FWER (confirmatoire) ou FDR (criblage).
2. Choisir la statistique (studentisée de préférence) et la façon d'obtenir les valeurs p (loi

théorique, permutation si échangeabilité, bootstrap sinon) ; s'assurer que $B \gg m/\alpha$ si on utilise des valeurs p ré-échantillonnées.

3. Examiner l'histogramme des valeurs p brutes.
4. Choisir la procédure selon la dépendance plausible : Holm ou maxT/minP (FWER), BH ou BY (FDR).
5. Rapporter les valeurs p ajustées, le nombre de tests et la procédure, et valider les découvertes.

Exercices

Exercice 8.1 (Résolution de problèmes, tests multiples, H26)**.** Dans une expérience, on teste 20 hypothèses nulles $H_0(j) : \mu_j \leq 0$ contre $H_1(j) : \mu_j > 0$, $j = 1, \ldots, 20$, avec des statistiques $Z(j)$ de loi $\mathcal{N}(0,1)$ sous H_0, supposées indépendantes. Les valeurs p unilatérales, en ordre croissant, sont :

j	1	2	3	4	5	6	7	8	9	10
$p_{(j)}$	0,0004	0,0005	0,0008	0,0017	0,0018	0,0019	0,0024	0,0024	0,0025	0,0034
j	11	12	13	14	15	16	17	18	19	20
$p_{(j)}$	0,0042	0,0046	0,0046	0,0052	0,0088	0,0492	0,0740	0,4831	0,6186	0,6364

a) Calculez les suites $0{,}05/(21 - j)$ et $0{,}05\, j/20$.

b) Selon la procédure descendante de Holm, quelles hypothèses sont rejetées à un FWER de 0,05 ? Expliquez votre démarche.

c) Selon la procédure ascendante de Benjamini et Hochberg, quelles hypothèses sont rejetées à un FDR de 0,05 ? Expliquez votre démarche.

d) Expliquez pourquoi les procédures de contrôle du FDR rejettent plus d'hypothèses nulles que celles du FWER au même niveau.

Exercice 8.2 (Examen final 2023, question 4)**.** On veut découvrir lesquelles de $m = 200$ substances potentiellement toxiques présentes dans des usines sont associées à une maladie, en les testant une à une par un modèle de régression. Il s'agit d'un criblage initial ; la toxicité des substances détectées sera validée expérimentalement par la suite.

a) Proposez un critère de contrôle de l'erreur de 1re espèce sur ces 200 tests et justifiez-le.

b) Supposons que les statistiques de Wald des coefficients suivent une loi normale et que la corrélation entre les statistiques des différentes substances soit nulle ou positive. Proposez une procédure pour contrôler l'erreur selon le critère choisi en a), en la décrivant telle qu'elle s'applique à ces statistiques.

Exercice 8.3 (Questions de pratique pour l'examen final, H26, question 3)**.** Dans une enquête de couverture vaccinale, on veut tester l'association entre le statut vaccinal et chacun des 20 prédicteurs sociodémographiques inclus dans un modèle de régression logistique. On s'attend à ce que certains prédicteurs soient corrélés.

a) Proposez un critère de contrôle de l'erreur de 1re espèce sur ces 20 tests et justifiez-le.

b) Proposez une procédure pour contrôler l'erreur selon ce critère, en décrivant et en justifiant chaque étape telle qu'elle s'applique aux statistiques de Wald. Sous quelle condition supplémentaire pourrait-on utiliser une procédure ascendante ?

Exercice 8.4 (Variante de l'examen final H26, question 3)**.** On teste $m = 5000$ hypothèses, dont 4900 sont vraies. La puissance du test pour une hypothèse fausse est 0,95 au seuil 0,05, 0,70 au seuil qui contrôle le FDR à 0,10, et 0,50 au seuil de Bonferroni pour un FWER de 0,05. Calculez $\mathrm{E}[R]$ et $\mathrm{E}[V]$ dans les trois cas, en approximant le FDR par $\mathrm{E}[V]/\mathrm{E}[R]$. Quel seuil individuel la procédure BH utilise-t-elle approximativement ?

Exercice 8.5 (Propriétés des taux et des procédures)**.** a) Montrez que, pour tout jeu de valeurs p, la procédure de Hochberg rejette au moins autant d'hypothèses que celle de Holm, et que BH rejette au moins autant que Hochberg.

b) Vérifiez la formule de la valeur p ajustée de la procédure descendante de Šidák et montrez qu'elle est toujours inférieure ou égale à celle de Holm.

c) (Doctorat) Montrez que la procédure fermée fondée sur des tests locaux de Bonferroni est la procédure de Holm.

Exercice 8.6 (Doctorat)**.** Montrez que BH revient à rejeter $H_0(j)$ si $p_j \leq \hat{t}$, où $\hat{t} = \max\{t \in \{p_{(1)}, \ldots, p_{(m)}\} : mt/R(t) \leq \alpha\}$ et $R(t) = \#\{j : p_j \leq t\}$. Proposez une version adaptative en remplaçant m par $\hat{m}_0 = m\,\hat{\pi}_0(\lambda)$; pourquoi utiliser $\#\{p_j > \lambda\} + 1$ en petit échantillon ?

Références

— Alon, U., Barkai, N., Notterman, D. A. et coll. (1999). Broad patterns of gene expression. *PNAS*, 96, 6745–6750.

— Benjamini, Y. et Hochberg, Y. (1995). Controlling the false discovery rate. *JRSS B*, 57, 289–300.

— Benjamini, Y. et Hochberg, Y. (2000). On the adaptive control of the false discovery rate. *J. Educ. Behav. Stat.*, 25, 60–83.

— Benjamini, Y., Krieger, A. M. et Yekutieli, D. (2006). Adaptive linear step-up procedures. *Biometrika*, 93, 491–507.

— Benjamini, Y. et Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29, 1165–1188.

— Dudoit, S. et van der Laan, M. J. (2008). *Multiple Testing Procedures with Applications to Genomics*. Springer.

— Farcomeni, A. (2010). Multiple testing methods. Ch. 3 dans F. Emmert-Streib et M. Dehmer (éd.), *Medical Biostatistics for Complex Diseases*. Wiley-VCH.

— Ge, Y., Dudoit, S. et Speed, T. P. (2003). Resampling-based multiple testing for microarray data. *Test*, 12, 1–77.

— Golub, T. R., Slonim, D. K., Tamayo, P. et coll. (1999). Molecular classification of cancer. *Science*, 286, 531–537.

— Hochberg, Y. (1988). A sharper Bonferroni procedure. *Biometrika*, 75, 800–802.

— Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, 6, 65–70.

— Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75, 383–386.

— Janssen, A. (1997). Studentized permutation tests for non-i.i.d. hypotheses. *Statist. Probab. Lett.*, 36, 9–21.

— Marcus, R., Peritz, E. et Gabriel, K. R. (1976). On closed testing procedures. *Biometrika*, 63, 655–660.

— Sarkar, S. K. (1998). Probability inequalities for ordered MTP_2 variables. *Ann. Statist.*, 26, 494–504.

— Šidák, Z. (1967). Rectangular confidence regions for multivariate normal means. *JASA*, 62, 626–633.

— Simes, R. J. (1986). An improved Bonferroni procedure. *Biometrika*, 73, 751–754.

— Storey, J. D. (2002). A direct approach to false discovery rates. *JRSS B*, 64, 479–498.

— van der Laan, M. J., Dudoit, S. et Pollard, K. S. (2004). Augmentation procedures. *Stat. Appl. Genet. Mol. Biol.*, 3, article 15.

— Westfall, P. H. et Young, S. S. (1993). *Resampling-Based Multiple Testing*. Wiley.

— Yekutieli, D. et Benjamini, Y. (1999). Resampling-based FDR controlling procedures for correlated statistics. *J. Statist. Plann. Inference*, 82, 171–196.