Tests multiples

1. Introduction

On mesure le niveau d'expression d'un ensemble de gènes $i = 1, \dots, m$ au sein de 2 populations (tissus sains, tissus tumoraux) de tailles n_1 et n_2. On pose $n = n_1 + n_2$. On se place dans le cadre où m est grand devant n. On cherche à savoir quels sont les gènes différentiellement exprimés dans les deux populations.

On note $X = (X^1, \dots, X^n) = (Y^1, \dots, Y^{n_1}, Z^1, \dots, Z^{n_2}) \in \mathbb{R}^{m \times n}$

et on suppose que $Y^1, \dots, Y^{n_1}$ i.i.d de même loi que Y $(\in \mathbb{R}^m)$

$Z^1, \dots, Z^{n_2}$ ——— Z $(\in \mathbb{R}^m)$

On note P la loi de X (v.a. sur $\mathbb{R}^{m \times n}$) et $\mathcal{P}$ la famille à laquelle appartient P et

$$\mu_{i,1}(P) = \mathbb{E}(Y_i) \quad \text{où } Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix}$$

$$\mu_{i,2}(P) = \mathbb{E}(Z_i) \quad Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_m \end{pmatrix}$$

On veut décider si $P \in \Theta_{0,i}$ où $\Theta_{0,i} = \{P \in \mathcal{P}, \mu_{i,1}(P) = \mu_{i,2}(P)\}$

Pour tout $i = 1, \dots, m$ on considère des hypothèses

$$H_{0,i} : \mu_{i,1}(P) = \mu_{i,2}(P) \quad \text{v.s.} \quad H_{1,i} = \overline{H_{0,i}}$$

On note $\mathcal{H}_0(P) = \{i = 1, \dots, m, \ H_{0,i} \text{ est vérifiée}\}$

= ensembles des hypothèses nulles vérifiées

(= ensemble des gènes non-différentiellement exprimés)

et $\mathcal{H}_1(P) = \{i = 1, \dots, m, \ H_{0,i} \text{ n'est pas vérifiée}\}$

$= \{1, \dots, m\} \setminus \mathcal{H}_0(P)$.

(= ensemble des gènes différentiellement exprimés)

On commence par construire les tests individuels, pour chaque hypothèse $H_{0,i}$. Dans le cas considéré, dans ce paragraphe, on va construire des tests de Student.

On pose $\overline{Y_i} = \frac{1}{n_1} \sum_{j=1}^{n_1} Y_i^j$ et $\overline{Z_i} = \frac{1}{n_2} \sum_{j=1}^{n_2} Z_i^j$

EXO

En supposant que $Y_i^j \sim \mathcal{N}(\mu_{i,1}, \sigma^2)$ et $Z_i^j \sim \mathcal{N}(\mu_{i,2}, \sigma^2)$ et que $(Y^1, \dots, Y^{n_1}) \perp\!\!\!\perp (Z^1, \dots, Z^{n_2})$, montrer que

$$\sqrt{\frac{n_1 n_2}{n \hat{\sigma}^2}} \left(\overline{Y_i} - \overline{Z_i}\right) \sim St(n-2) \quad \text{si } \mathcal{H}_{0,i} \text{ est vérifiée.}$$

où $\hat{\sigma}^2 = \frac{1}{n-2} (S_i^1 + S_i^2)$ et $S_i^1 = \sum_{j=1}^{n_1} (Y_i^j - \overline{Y_i})^2$

$S_i^2 = \sum_{j=2}^{n_2} (Z_i^j - \overline{Z_i})^2$

On choisit donc la statistique de test

$$\mathcal{T}_i(x) = \sqrt{\frac{n_1 n_2}{n \hat{\sigma}^2}} \left|\overline{Y_i} - \overline{Z_i}\right| \quad \text{et on rejette } \mathcal{H}_{0,i}$$

lorsque $\mathcal{T}_i(x) > s_\alpha$ (où s_α est un seuil à fixer suivant la valeur de α)

Définissons $T_{P,i}(s) = \mathbb{P}_{X \sim P}(\mathcal{T}_i(x) \geq s)$. Si $\mathcal{H}_{0,i}$ est vérifiée (de façon équivalente si $P \in \Theta_{0,i}$) on a

$$\begin{aligned} T_{P,i}(s) &= \mathbb{P}(|\tau_{n-2}| \geq s) \quad \text{où } \tau_{n-2} \sim St(n-2) \\ &= 2\mathbb{P}(\tau_{n-2} \geq s) \quad \text{par symétrie de la } St(n-2) \\ &= 2(1 - F_{St(n-2)}(s)) \quad \text{où } F_{St(n-2)} \text{ est la f.d.r de la loi } St(n-2) \end{aligned}$$

<u>Définition</u> : p-value.

On définit la p-value $p_i(x)$ de ce test comme

$$p_i(x) = \sup_{P \in \Theta_{0,i}} T_{P,i}(\mathcal{T}_i(x))$$

Dans notre exemple, $\forall P \in \Theta_{0,i}$ $T_{P,i}(\mathcal{T}_i(x)) = 2(1 - F_{St(n-2)}(\mathcal{T}_i(x)))$

donc $p_i(x) = 2(1 - F_{St(n-2)}(\mathcal{T}_i(x)))$ (le sup n'est pas nécessaire dans cet exemple)

EXO

Montrons que dans notre exemple $T_{P,i}(\mathcal{T}_i(x)) = p_i(x)$ suit une loi uniforme sur $[0,1]$ quand $P \in \Theta_{0,i}$

~~$\mathbb{P}(T_{P,i}$~~ $\mathbb{P}(p_i(x) \leq u) = \mathbb{P}(T_{P,i}(\mathcal{T}_i(X)) \leq u)$

$= \mathbb{P}(2(1 - F_{St(n-2)}(\mathcal{T}_i(X))) \leq u) = \mathbb{P}(1 - \frac{u}{2} \leq F_{St(n-2)}(\mathcal{T}_i(X)))$

$$= \mathbb{P}\left(F^{-1}_{St(n-2)}\left(1-\frac{u}{2}\right) \leq T_i(X)\right)$$
$$= 2\left(1 - F_{St(n-2)}\left(F^{-1}_{St(n-2)}\left(1-\frac{u}{2}\right)\right)\right)$$ ← car $P \in \Theta_{0i}$ donc $T_i(X) \sim St(n-2)$
$$= u. \quad \blacksquare$$

Cette propriété est très souvent vérifiée et est fondamentale pour les tests multiples.

On passe maintenant aux tests multiples. On aura donc un vecteur $(p_i(X))_{i=1,\dots,m}$ de p-values et on va rejeter les hypothèses $\mathcal{H}_{0,i}$ pour lesquelles les p-values associées sont "petites" (← à définir).

Avant cela, on va écrire sous forme de théorème la propriété de l'exercice précédent.

2. Propriété fondamentale des p-values

On considère dans ce paragraphe que $m = 1$.

On adapte donc les notations $\mathcal{H} : P \in \Theta_0$, la statistique de test est $T(X)$, la zone de rejet $\{T(X) \geq s\}$, $T_P(s) = \mathbb{P}_{X \sim P}(T(X) \geq s)$.

ici $T(X)$ est générale, pas forcément la stat. d'un test de Student

On note de plus $F_P(s) = \mathbb{P}_{X \sim P}(T(X) \leq s)$

$$\mathring{F}_P(s) = \mathbb{P}_{X \sim P}(T(X) < s)$$

(remarquez qu'on a $\mathring{F}_P(s) \leq F_P(s)$ $\forall s \in \mathbb{R}$, avec égalité fonctionnelle si la loi admet une densité).

On peut alors écrire :

$$T_P(s) = \mathbb{P}_{X \sim P}(T(X) \geq s) = 1 - \mathbb{P}_{X \sim P}(T(X) < s)$$
$$= 1 - \mathring{F}_P(s).$$

Propriété

La p-value définie par $p(X) = \sup_{P \in \Theta_0} T_P(T(X))$ vérifie.

i) $p(X)$ est stochastiquement majorée par la loi $\mathcal{U}_{[0,1]}$ ($\forall P \in \Theta_0$ $\forall u \in [0,1]$ $\mathbb{P}(p(X) \leq u) \leq u$.

ii) Si F_P est continue pour tout $P \in \Theta_0$ alors
$p(x) = \min\{\alpha \in [0,1],\ T(x) \geq \sup_{P \in \Theta_0} F_P^{-1}(1-\alpha)\}$
et si Θ_0 est un singleton $p(X) \sim \mathcal{U}_{[0,1]}$

Preuve

Point i) $\forall P \in \Theta_0$ et $\alpha \in [0,1]$.

$\{T_P(\mathcal{T}(x) \leq \alpha\} = \{1 - \mathring{F}_P(\mathcal{T}(x)) \leq \alpha\} = \{1-\alpha \leq \mathring{F}_P(\mathcal{T}(x))\}$. par définition

On veut montrer que $\{T_P(\mathcal{T}(X)) \leq \alpha\} \subset \{\mathcal{T}(x) \geq F_P^{-1}(1-\alpha)\}$. En effet. si c'est vérifié on a alors.

$$\begin{aligned}\mathbb{P}(p(X) \leq \alpha) &= \mathbb{P}\Big(\sup_{P\in\Theta_0} T_P(\mathcal{T}(x)) \leq \alpha\Big) \\ &\leq \mathbb{P}_{X\sim P\in\Theta_0}\big(T_P(\mathcal{T}(x)) \leq \alpha\big) \\ &\leq \mathbb{P}_{X\sim P\in\Theta_0}\big(\mathcal{T}(x) \geq F_P^{-1}(1-\alpha)\big) \\ &\leq \mathbb{P}_{X\sim P\in\Theta_0}\big(\mathcal{T}(x) > F_P^{-1}(1-\alpha)\big) = 1 - F_P(F_P^{-1}(1-\alpha)) = \alpha.\end{aligned}$$

car sup.

grâce à l'inclusion

où $F_P^{-1}(u) = \inf(x \in \mathbb{R},\ F_P(x) \geq u)$. (c'est une inverse généralisée).

Montrons donc l'inclusion

$$\{T_P(\mathcal{T}(X)) \leq \alpha\} = \{1 - T_P(\mathcal{T}(X)) \geq 1-\alpha\} = \{\mathring{F}_P(\mathcal{T}(x)) \geq 1-\alpha\}.$$

$$\subset \{F_P(\mathcal{T}(x)) \geq 1-\alpha\} = \{\mathcal{T}(x) \geq F_P^{-1}(1-\alpha)\} \quad \square.$$

car $\mathring{F}_P \leq F_P$; par définition de F_P^{-1}

Point ii).

Si tous les $P \in \Theta_0$ sont absolument continues $\mathring{F}_P = F_P$ l'inclusion précédente est donc une égalité.

Comme $p(x) = \sup_{P\in\Theta_0} T_P(\mathcal{T}(x))$ on peut également la définir

$$\begin{aligned}\text{comme } p(x) &= \min\{\alpha \in [0,1],\ \forall P \in \Theta_0,\ T_P(\mathcal{T}(x)) \leq \alpha\} \\ &= \min\{\alpha \in [0,1],\ \forall P \in \Theta_0,\ \mathcal{T}(x) \geq F_P^{-1}(1-\alpha)\} \\ &= \min\{\alpha \in [0,1],\ \mathcal{T}(x) \geq \sup_{P\in\Theta_0} F_P^{-1}(1-\alpha)\}.\end{aligned}$$

grâce au fait que l'inclusion est une égalité dans ce cas.

Si Θ_0 est un singleton, on a

$$\begin{aligned}\mathbb{P}(p(x) \leq u) &= \cancel{\mathbb{P}(\min\{\alpha \in [0,1],\ \mathcal{T}(x) \geq F_P^{-1}(1-\alpha)\} \leq u)} \\ &= \cancel{\mathbb{P}(\min\{\alpha \in [0,1],\ F_P(\mathcal{T}(x)) \geq 1-\alpha\} \leq u)} \\ &= \cancel{\mathbb{P}(F_P^{-1}(F_P(1-\alpha)) \leq u)} \\ &= \mathbb{P}\Big(\sup_{P\in\Theta_0} T_P(\mathcal{T}(x)) \leq u\Big) = \mathbb{P}_{X\sim P}(T_P(\mathcal{T}(x)) \leq u) \\ &= \dots = u.\end{aligned}$$

donc $p(X) \sim \mathcal{U}_{[0,1]}$ $\square$.

3 Cadre général des tests multiples. $(m > 1)$

On note $\#(\mathcal{H}_0(P)) = m_0(P)$ nombre d'hypothèses nulles vérifiées par P. A partir de l'observation de X on veut découvrir $\mathcal{H}_1(P)$.

Note $\#(A) = $ cardinal(A)

On va donc utiliser le vecteur $p = (p_i(X))_{i=1,\dots,m} \in [0,1]^m$.
avec $\forall P \in \mathcal{P}\ \forall i \in \mathcal{H}_0(P)\ \forall u \in]0,1]\ \mathbb{P}(p_i(X) \leq u) \leq u$.
On note le vecteur $(p_i(X))_{i=1,\dots,m}$ quand la dépendance en X n'est pas importante.

[Définition : une procédure de tests multiples est une fonction $R : q = (q_i)_{i=1,\dots,m} \in [0,1]^m \to R(q) \in \{1,\dots,m\}$

L'ensemble $R(p)$ est donc l'ensemble des hypothèses nulles $\mathcal{H}_{0,i}$ rejetées par la procédure.

On considère dans la suite des procédures particulières, les procédures de seuillage qui prennent la forme.

$$R(q) = \{ i = 1,\dots,m \ ,\ q_i \leq \hat{t}(q) \}$$

<u>Exemple</u> procédure de Bonferroni (de niveau α).

$$R(q) = \{ i = 1,\dots,m \ ,\ q_i \leq \frac{\alpha}{m} \}.$$

4 <u>Erreurs de type I</u>

On conserve l'approche de Neyman-Pearson qui consiste à contrôler l'erreur de type I en essayant de rendre celle de type II la plus petite possible.

Dans un cadre multi-tests, il y a plusieurs notions d'erreurs de type I.

a) k-FWER : k-family error rate.

$$k\text{-FWER}(R(p), P) = \mathbb{P}\left(\#(R(p) \cap \mathcal{H}_0(P)) \geq k \right)$$

= proba que R fasse plus de k faux rejets quand la loi est P !

b). FDP = false discovery proportion.

$$\text{FDP}(R(p), P) = \frac{\#(R(p) \cap \mathcal{H}_0(P))}{\#(R(p)) \vee 1} \quad \leftarrow \text{car } \#(R(p)) \text{ peut être nul!}$$

= proportion d'erreurs dans l'ensemble des hypothèses rejetées.

⚠ la FDP est une quantité aléatoire : elle dépend de $p = p(X)$

c) FDR = false discovery rate

$$\text{FDR}(R(p), P) = \mathbb{E}(\text{FDP}(R(p), P))$$

Exemple procédure de Bonferroni $R^{Bonf}(p) = \{i=1,\dots,m,\ p_i = p_i(x) \leq \frac{\alpha}{m}$ 6

Dans ce cas $\#(R(p) \cap \mathcal{H}_0(P)) = \#\{i \in \mathcal{H}_0(P),\ p_i(X) \leq \frac{\alpha}{m}\}$

On a donc

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

par minoration par 1 du dénominateur

$$\leq \mathbb{E}(\#(R(p) \cap \mathcal{H}_0(P)))$$

$$= \sum_{i \in \mathcal{H}_0(P)} \mathbb{P}\left(p_i(X) \leq \frac{\alpha}{m}\right)$$

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

Exemple : procédure de Holm.

A partir du vecteur des p-values $(p_i)_{i=1,\dots,m}$, on crée celui des p-values ordonnées $p(1) \leq \dots \leq p(m)$ auquel on associe les hypothèses correspondantes $\mathcal{H}_{0,(1)} \dots \mathcal{H}_{0,(m)}$

On définit l'indice minimale k pour lequel $p(k) \geq \frac{\alpha}{m+1-k}$.

On rejette les hypothèses $\mathcal{H}_{0,(1)}, \dots, \mathcal{H}_{0,(k-1)}$ (et on accepte $\mathcal{H}_{0,(k)}, \dots, \mathcal{H}_{0,(m)}$).

On note k^0 la première hypothèse de $\mathcal{H}_0(P)$ rejetée. Avec ces définitions, les hypothèses $\mathcal{H}_{0,(1)}, \dots, \mathcal{H}_{0,(k^0-1)}$ sont toutes rejetées à raison. On a donc $k^0 - 1 \leq m - m_0(P)$

$$\Leftrightarrow \frac{1}{m-k^0+1} \leq \frac{1}{m_0(P)} \quad \text{et} \quad p(k^0) \leq \frac{\alpha}{m+1-k^0} \leq \frac{\alpha}{m_0(P)}$$

car (k^0) est rejetée.

On peut alors calculer :

$$\mathbb{P}(\#(R(p) \cap \mathcal{H}_0(P)) \geq 1) = \mathbb{P}\left(\exists i \in \mathcal{H}_0(P)\ p_i \leq \frac{\alpha}{m_0(P)}\right) \leq \alpha.$$

un peu rapide, réécrire les dé tout

Donc la procédure de Holm contrôle le 1-FWER.