Points importants du cours 4

Vincent Guillemot

28 janvier 2009

1 Tests d'hypothèses

1.1 Exemple préliminaire

(cf. cours précédent)

1.2 Démarche générale

(cf. cours précédent)

1.3 Test paramétrique

(cf. cours précédent)

1.4 Tests non paramétriques

1.4.1 Notion et définitions

Un test dit non paramétrique permet de tester la validité d'une hypothèse sans avoir à formuler d'hypothèse supplémentaire sur les lois des variables aléatoires étudiées. Pour la suite, X et Y sont des v.a.r. indépendantes, $(X_i)_{i\in[[1;n_x]]}$ un échantillon de taille n_x relativement à X, $(Y_i)_{i\in[[1;n_y]]}$ un échantillon échantillon de taille n_y relativement à Y.

Définition 1 - *Rang*

Le rang de X_i dans l'échantillon $X_1, ..., X_n$ est la « position » de la variable X_i dans la liste de ces variables réordonnées,

$$X_{(1)} < X_{(2)} < ... < X_{(n)}$$

Cette notion de rang est valable quand la variable étudiée est « continue ». Quand il y a des réalisation de rangs identiques (des **« ex-æquo »**), la stratégie la plus souvent adoptée est de leur attribuer la moyenne de leurs rangs s'ils n'avaient pas été ex-æquos.

Exemple : En présence d'ex-æquos, la stratégie est la suivante :

x_i	0.4	-0.5	-1.4	-0.5	-1.3
Rang	5	?	1	?	2
Rang ajusté	5	3.5	1	3.5	2

Les tests non paramétriques que nous allons voir sont les équivalents des tests de Student pour la comparaison de deux moyennes empiriques. Attention toutefois aux noms des tests présentés, ils peuvent porter un nom différent d'un ouvrage à l'autre et surtout d'un **logiciel à l'autre**.

1.4.2 Test de Wilcoxon (sur des données appariées)

Ce test est l'équivalent du test de Student sur des échantillons appariées : on travaillera donc sur la variable $D = X - Y$. L'hypothèse nulle peut être formulée de différentes manières, pour ce genre de tests. Nous choisissons la formulation de Saporta : $\mathcal{H}_0 =$ « la médiane de D nulle » contre $\mathcal{H}_1 =$ « la médiane de D est décalée ».

Pour des données appariées, on peut poser $n_x = n_y = n$. L'échantillon associé à D est noté $D_1, ..., D_n$.

Les tests non paramétriques ont des fonctionnement généralement séquentiels très facilement transposables en un algorithme :

1. Considérer les différences entre variables : $D_i = X_i - Y_i$.
2. Ranger ces D_i par ordre croissant de valeur absolue.
3. Faire la somme des rangs correspondants à $D_i > 0$.

On note cette statistique W_+.

Théorème 1 - *Moments de la statistique W_+*

$$E(W_+) = \frac{n(n+1)}{4} \text{ et } V(W_+) = \frac{n(n+1)(2n+1)}{24}.$$

Quand $n \geq 10$, on peut approximer cette variable par une variable aléatoire gaussienne. Ainsi, la décision peut se formuler ainsi :

$$\frac{|W_+ - E(W_+)|}{\sqrt{V(W_+)}} \geq u_{1-\alpha/2} \Rightarrow \mathcal{RH}_0$$

1.5 Test de Mann-Whitney (données non appariée)

Dans le cas de données non appariées, $\mathcal{H}_0 =$ « X et Y ont même distribution » et $\mathcal{H}_1 =$ « X et Y ont des distributions différentes ». Ces hypothèses semblent bien sûr très fortes. En fait, le test de Mann-Whitney teste surtout si les deux échantillons ont des médianes égales, et donc des moyennes égales si les variables considérées ont des lois symétriques.

La procédure à suivre est la suivante :

1. Rassembler les observations X_i et Y_j : on obtient au total $n_x + n_y$ variables.
2. Les classer.
3. Calculer la somme des rangs attribués aux X_i.

Cette statistique est notée W_X, c'est une statistique de Wilcoxon comme la précédente.

On peut également compter le nombre total de couples (X_i, Y_j) tels que X_i a un rang plus grand que Y_j. On note cette statistique U. Cette dernière est la statistique de Mann-Whitney.

Théorème 2 - *Lien entre U et W_X*

Moments de U :

$$E(U) = \frac{n_x n_y}{2} \text{ et } V(U) = \frac{n_x n_y (n_x + n_y + 1)}{12}.$$

Moments de W_X :

$$E(W_X) = \frac{n_x(n_x + n_y + 1)}{2} \text{ et } V(W_X) = \frac{n_x n_y (n_x + n_y + 1)}{12}.$$

Ces deux statistiques respectent la relation suivante :

$$W_X = n_x n_y + \frac{n_x(n_x + 1)}{2} - U.$$

On décide de rejeter $\mathcal{H}_0$ comme suit :

$$|W_X - \frac{n_x(n_x + n_y + 1)}{12}| > u_{\alpha/2}\sqrt{\frac{n_x n_y (n_x + n_y + 1)}{12}} \Rightarrow \mathcal{R}\mathcal{H}_0.$$

1.6 Tests multiples

Lorsque l'on mène en parallèle beaucoup de tests de la même nature, il est dangereux d'utiliser la p-value pour sélectionner les conditions (les variables) pour lesquelles $\mathcal{H}_0$ est rejetée.

Par exemple : sur des données transcriptomiques, on réalise très souvent une analyse différentielle c'est à dire une identification des gènes différentiellement exprimés entre deux situations biologiques différentes.

Problème, on effectue très souvent plusieurs milliers de tests !

Définition 2 - *Faux positif*

On dit qu'un gène est un faux positif si le test mené pour déterminer s'il est DE amène à rejeter H0 alors qu'en vérité on il respecte H0.

Tableau des faux positifs pour un seuil, par exemple t, donné et m tests effectués :

	$\mathcal{H}0$	$\mathcal{H}_1$	total
Rejetés	$FP(t)$	$VP(t)$	$R(t)$
Acceptés	$VN(t)$	$FN(t)$	$m - R(t)$
	m_0	$m_1 = m - m_0$	

Exemple : Soit une puce à 10 000 gènes dont 100 sont différentiellement exprimés : $\mathcal{H}_0$ est vraie dans 9 900 cas. Considérons $P(\mathcal{H}_1|\mathcal{H}_0) = 5\% = \alpha$. Le nombre de faux positifs est alors $\alpha * 9900 = 495$. Supposons que le test soit puissant et que $\beta \approx 0$. Dans ce cas, $495 + 100 = 595$ gènes seront déclarés différentiellement exprimés, soit un taux de faux positifs égal à $495/595 = 0.83$!!

Définition 3 - *FWER*

Family Wise Error Rate : $FWER(t) = P(FP(t) \geq 1)$

Définition 4 - *FDR*

False Discovery Rate : Si $R(t) = 0$: $FDR = 0$, sinon $FDR = E(VP(t)/R(t))$

Le FWER contrôle la probabilité de ne n'avoir aucun faux positif, le FDR contrôle la proportion de faux positifs. Le FWER ne repère que très peu de vrais positifs, il est très conservatif et peu adapté aux puces. Dans le contexte de l'analyse différentielle, le FDR est préférable car il permet de :

- détecter davantage de vrais positifs,
- contrôler la proportion de faux positifs.

La méthode de *Bonferroni* permet un contrôle du FWER La méthode de *Benjamini-Hochberg* permet un contrôle du FDR.

Exercice sous R : Simuler un jeu de données avec 1000 gènes dont 100 sont différentiellement exprimés. Faire un t-test pour identifier les gènes différentiellement exprimés :

- sans contrôle sur la p-value
- avec un contrôle sur la p-value grâce à la méthode `p.adjust`