Introduction aux Statistiques non paramétriques

d'après Eric Wajnberg (INRA-INRIA)
Université de Nice-Sophia-Antipolis (modifié).

Préambule

Lorsque les conditions de réalisation des tests du modèle linéaire général ou du modèle linéaire généralisé ne sont pas vérifiées (distribution des variables non identifiées ou non identifiables, non-égalité des variances, effectifs petits, etc.), il convient d'utiliser d'autres méthodes qui permettent de s'affranchir de ces conditions. Il s'agit de méthodes dites non-paramétriques (en anglais « non-parametric », ou « distribution-free »), valides quelle que soit la distribution des données (y compris si les données sont Gaussiennes, Poissoniènnes, Binomiales ou inconnues, etc.). Cependant, la puissance des tests non-paramétriques (c'est-à-dire leur capacité à détecter un effet sur un même jeu de données) est généralement plus faible que leurs équivalents paramétriques.

Le fondement de ces tests est de remplacer les observations par leurs rangs, après les avoir classés par ordre croissant. C'est sur ces rangs et leurs moyennes, variances etc. que se feront les calculs.

Méthode :

Lorsque ce classement aboutit à des ex-aequos, leur rang moyen leur est attribué. Il ne sera fait aucun calcul sur les données elles-mêmes.

Par exemple, la suite de données suivante :
10, 20, 13, 45, 0, 12 (*classement* : 0, 10, 12, 13, 20, 45)
sera remplacée par :
2, 5, 4, 6, 1, 3

Et la suite suivante :
10, 20, 13, 45, 0, 13 (classement : 0,10, 13, 13, 20, 45)
sera remplacée par :
2, 5, 3.5, 6, 1, 3.5
(car la valeur « 13 » a pour rang aussi bien « 3 » que « 4 », et donc pour rang moyen $\frac{3+4}{2} = 3.5$).

Sous R, deux fonctions sont utiles pour cela :

```
sort(x) #fonction de tri du vecteur x
rank(x) #fonction d'attribution des rangs
```

Comparaison des moyennes de deux échantillons indépendants : Test de Wilcoxon - Mann-Whitney (parfois appelé test *U* de Mann-Whitney).

Soit à comparer les moyennes de deux échantillons. Le premier a un effectif de n_1, le second un effectif de n_2. La démarche est la suivante :

1. On classe tout d'abord les $n_1 + n_2$ valeurs par ordre croissant, et on les remplace par leur rang (ou rang-moyen en cas d'ex-aequos).
2. On calcule ensuite la somme «*S*» des rangs des valeurs issues du premier échantillon. Si la moyenne de ce premier échantillon est inférieure à celle du second, cette somme sera particulièrement faible. Dans le cas contraire, elle sera particulièrement élevée. C'est ce que nous allons tester.

3. On calcule ensuite :

$$E = \frac{n_1(n_1 + n_2 + 1)}{2}$$

C'est ce que devrait valoir *S*, si les deux moyennes sont égales, c'est-à-dire si l'hypothèse H_0 est vraie : il n'existe pas de différence significative entre les deux moyennes des échantillons.

4. On calcule également :

$$V = \frac{n_1 n_2(n_1 + n_2 + 1)}{12}$$

Ce qui correspond à la variance de la somme des rangs calculée.

5. Il ne reste plus qu'à calculer la quantité :

$$\varepsilon = \frac{|S - E|}{\sqrt{V}}$$

Si cette quantité est supérieure à 1.96 on rejette l'hypothèse d'égalité des moyennes des deux échantillons. Dans le cas contraire, on ne peut rejeter l'hypothèse.

Remarque : ce test n'est valide que si les effectifs n_1 et n_2 sont tous deux supérieurs à 8 !

Avec R, le calcul peut se faire selon deux syntaxes différentes, dépendant de la manière dont les données sont présentées :

```
wilcox.test(x1,x2)
wilcox.test(x~y)
```

Comparaison des moyennes de deux échantillons appariés : Test des rangs signés de Wilcoxon (*signed rank test*)

Nous voulons à présent comparer la moyenne de deux échantillons de valeurs appariées. La démarche la suivante est adoptée :

1. On calcule toutes les différences d_i entre le premier et le deuxième échantillon, et on classe leurs valeurs absolues.
2. Supposons qu'il y ait *n* différences non nulles.
3. On calcule alors la somme S des rangs des **valeurs positives**.
4. On calcule ensuite :

$$E = \frac{n(n + 1)}{4}$$

(C'est ce que devrait valoir S, si l'identité des deux populations est vraie)

5. On calcule également la variance de la somme des rangs :

$$V = \frac{n(n + 1)(2n + 1)}{24}$$

6. Il ne reste plus qu'à calculer la quantité :

$$\varepsilon = \frac{|S - E|}{\sqrt{V}}$$

Si cette quantité est supérieure à 1.96 on rejette l'hypothèse d'égalité des moyennes des deux échantillons. Dans le cas contraire, on ne peut rejeter l'hypothèse.

Remarque : ce test n'est valide que si n >25 !

Avec R, le calcul peut se faire selon deux syntaxes différentes selon la manière dont les données sont présentées :

```
wilcox.test(x1,x2,paired=TRUE)
wilcox.test(x~y,paired=TRUE)
```

Ou bien encore, ce qui est évidement équivalent :

```
wilcox.test(x1-x2)
```

Comparaison de plusieurs moyennes (ANOVA non-paramétrique) : Test de Kruskal et Wallis

Nous avons cette fois-ci k échantillons d'effectifs n_1, n_2, ..., n_k, dont nous voulons comparer les moyennes. La démarche suivante est adoptée :

1. On classe tout d'abord les $n_1 + n_2 + ... + n_k$ valeurs par ordre croissant, et on les remplace par leur rang (ou rang-moyen en cas d'ex-aequos).
2. On calcule ensuite les sommes S_i des rangs des valeurs des k échantillons.
3. En notant $n = n_1 + n_2 + ... + n_k$, on calcule la quantité suivante :

$$\frac{12}{n(n+1)}\left\{\sum_{i=1}^{k}\frac{S_i^2}{n_i}\right\} - 3(n+1)$$

que l'on compare à la valeur lue dans une table de χ^2 à k-1 degrés de liberté, et au risque souhaité. Si la valeur calculée est supérieure à la valeur lue dans la table, on rejette l'hypothèse H_0 d'égalité des moyennes des k échantillons, au risque souhaité. Dans le cas contraire, on ne peut rejeter l'hypothèse.

Avec R, le calcul peut se faire selon la manière suivante :

```
kruskal.test(x~y)
```

Notons que l'on peut connaitre si besoin est, avec R, le risque associé à un χ^2 d'une valeur ch, à d degrés de liberté de la manière suivante :

```
1-pchisq(ch,d)
```

Test de corrélation entre deux variables : Test de Spearman

Pour déceler une liaison entre deux variables x et y de façon non-paramétrique, on procède de la manière suivante :

1. On classe les n valeurs de la variable x et on les remplace par leur rang (ou rang moyen en cas d'ex-aequos). Notons ces rangs S_i.
2. On classe les n valeurs de la variable y et on les remplace par leur rang (ou rang moyen en cas d'ex-aequos). Notons ces rangs R_i.

3. On calcule ensuite les différences $D_i = S_i - R_i$.
4. Puis la quantité :

$$R = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} D_i^2$$

La quantité $R\sqrt{n-1}$ suit une loi Normale Centrée Réduite, lorsque *n* est suffisamment grand.

Si la valeur obtenue pour cette quantité est supérieure à 1.96, on rejette l'hypothèse d'une absence de liaison entre les deux variables *x* et *y* avec 5% de risque. Sans le cas contraire, on ne peut rejeter l'hypothèse.

Remarque : ce test n'est valide que lorsque *n* est assez grand ($n \geq 8$ ou 10).

Avec R, le calcul peut se faire par exemple selon la manière suivante :

```
plot(x,y)
plot(rank(x),rank(y))
cor(x,y,method="spearman")
cor(rank(x),rank(y))
cor.test(x,y,method="spearman")
cor.test(rank(x),rank(y))
```