

Statistique non paramétrique

TP 3 : Tests de comparaison de K échantillons indépendants

<http://math.univ-lille1.fr/~jacques/enseignement.html>

Important : vous rédigerez un compte-rendu **soigneux** de votre travail en incluant notamment les scripts R et SAS utilisés.

1 Test de comparaison d'échantillons indépendants dans le cas paramétrique

Dans le cas paramétrique, on suppose que les échantillons sont de lois normales. Comparer des échantillons revient donc à comparer leur moyennes et leur variances. Dans le cas de 2 échantillons, on utilise les tests classiques de comparaison de moyenne et de variance de Student et Fisher. Dans le cas de plus de 2 échantillons, on effectue une analyse de variance à 1 facteur.

2 Test de comparaison de 2 échantillons indépendants : Wilcoxon-Mann-Whitney

Le test de Wilcoxon-Mann-Whitney permet de tester si deux échantillons indépendants sont issus d'une même population (homogénéité des échantillons indépendants).

Soient (X_1, \dots, X_n) et (X_1, \dots, X_m) deux échantillons indépendants issus de distributions continues. On se propose de tester :

$$H_0 : F_X = F_Y \quad \text{contre} \quad H_1 : F_X \neq F_Y$$

où F_X et F_Y sont les fonctions de répartition de X et de Y .

Le principe du test de Wilcoxon-Mann-Whitney consiste à ranger dans l'ordre croissant l'ensemble des observations mélangées des deux échantillons, à leur affecter un rang et à calculer séparément la somme des rangs des observations provenant de chacun des deux échantillons, notées Q_X et Q_Y .

En supposant que $n \leq m$, la statistique du test s'écrit :

$$U = Q_X - \frac{n(n+1)}{2}.$$

Sous H_0 , il existe des tables correspondant à cette statistique. Lorsque n et m sont tous les deux supérieurs à 8 (ce que l'on supposera dans ce TP), il est possible d'approximer la loi de cette statistique par une loi normale d'espérance $\mu = \frac{nm}{2}$ et de variance $\sigma^2 = \frac{nm(n+m+1)}{12}$.

Dans ce cas, le test revient donc à déterminer la statistique

$$Z = \frac{|U - \mu| - \frac{1}{2}}{\sigma}$$

et à la comparer avec les valeurs critiques d'une loi normale centrée réduite.

Remarque : la quantité $\frac{1}{2}$ correspond à la correction de continuité dans l'approximation de la statistique du test par une loi normale.

2.1 Mise en oeuvre sous SAS

Sous SAS la procédure `npar1way` réalise de test de Wilcoxon-Mann-Whitney si on lui précise l'option `wilcoxon` comme suit :

```
proc npar1way data = "your_data_file" wilcoxon ;
  class groupe ;
  var variable ;
run ;
```

La variable `groupe` est la variable permettant d'identifier les groupes.

2.2 Exercice

1. Deux groupes d'élèves ingénieurs GIS effectuent les 11 exercices d'un TP de Statistique. Les deux groupes vont au bout du TP, dans les temps suivants (minutes) :

groupe 1	4	5	7	8	9	10	11	12	13	17	40
groupe 2	3	6	14	15	16	18	19	20	21	25	35

La question que l'on se pose est la suivante : les deux groupes sont-ils homogènes d'un point de vue rapidité pour traiter ces exercices ?

2. Implémenter le test de Wilcoxon-Mann-Whitney sous R pour répondre à cette question.
3. Vérifier vos résultats à l'aide de SAS.
4. Aurait-on pu répondre à cette question en utilisant un test paramétrique ? Si oui, faites-le et comparez les résultats.

3 Test de comparaison de K échantillons indépendants : Kruskal-Wallis

Le test de Kruskal-Wallis compare les moyennes de plusieurs échantillons indépendants : c'est la version non-paramétrique de l'analyse de variance.

En R, la procédure `kruskal.test` permet de réaliser ce test, tandis que sous SAS il faut utiliser la même procédure `npar1way` que pour le test de Wilcoxon-Mann-Whitney en lui enlevant l'option `wilcoxon`.

3.1 Exercice à réaliser sous SAS

Le jeu de données `airquality` de R contient des données relatives à la qualité de l'air à New-York du 1er mai 1973 au 30 septembre 1973. La question que l'on se pose est la suivante :

Peut-on considérer que la quantité d'ozone présente dans l'air est distribuée de façon similaire pour les 5 mois de l'étude ?

1. Représenter graphiquement les données relatives à chaque moi de l'étude (diagramme, boîte à moustache...), et interpréter.
2. Que pouvez-vous dire graphiquement de l'égalité des moyennes, médianes et variances ?
3. Afin de répondre à la question, tester l'égalité des moyennes de chaque mois à l'aide du test de Kruskal-Wallis.
4. Si les 5 moyennes ne sont pas toutes égales, effectuer alors les tests 2 à 2.
5. Analyser la normalité des données pour chaque groupes. Aurais-t-on pu utiliser des tests paramétriques pour répondre à la question d'égalité des moyennes ? Si oui, faites-le.