Correction examen ModIA du 4 Février 2021

Cathy Maugis-Rabusseau et Jean-Yves TOURNERET

Exercice 1 : Comparaison des notes de deux groupes de TDs

Cet exercice s'inspire d'un exercice du livre de Gregory Corder et Dale Foreman intitulé "Nonparametric Statistics : A Step-By-Step Approach" (voir page 90). Un enseignant désire tester si ses deux groupes de TDs ont obtenus des résultats similaires lors de l'examen. Il a relevé les résultats suivants

Groupe 1	$x_1 = 16$	$x_2 = 13$	$x_3 = 16$	$x_4 = 16$	$x_5 = 13$	$x_6 = 9$	$x_7 = 12$	$x_8 = 12$	$x_9 = 20$	$x_{10} = 17$
Groupe 2	$y_1 = 11$	$y_2 = 2$	$y_3 = 10$	$y_4 = 4$	$y_5 = 9$	$y_6 = 8$	$y_7 = 5$	$y_8 = 6$	$y_9 = 4$	$y_{10} = 16$

1. Déterminer la valeur de la statistique du test de Mann-Whitney. Exprimer la p-valeur associée en fonction de la fonction de répartition de la loi normale $\mathcal{N}(0,1)$ notée F. Expliquer comment procéder pour décider si ces deux ensembles de résultats sont significativement différents avec un risque $\alpha = 0.05$.

   Le test de Mann-Whitney rejette l'hypothèse H_0 si $U \leq S_{1,\alpha}$ ou $U \geq S_{2,\alpha}$ avec $U = W - \frac{m(m+1)}{2}$, $W = \sum_{j=1}^{m} S_j$ et S_j est le rang de Y_j parmi les $n+m$ données réunies $(X_1, ..., X_n, Y_1, ..., Y_m)$. La $p-$valeur de ce test est définie par

$$\text{p-val} = 2\left[1 - F\left(\frac{|U - E[U]| - 0.5}{\sqrt{\text{var}[U]}}\right)\right].$$

   Dans cet exemple, on a les résultats suivants

   — Suite ordonnée

$$z(.) = (2, 4, 4, 5, 6, 8, 9, 9, 10, 11, 12, 12, 13, 13, 16, 16, 16, 16, 17, 20).$$

   — Données associées

$$y_2, y_4, y_9, y_7, y_8, y_6, y_5, x_6, y_3, y_1, x_7, x_8, x_2, x_5, x_1, x_3, x_4, y_{10}, x_{10}, x_9.$$

   — rangs du groupe 2

$$r_1 = 10, r_2 = 1, r_3 = 9, r_4 = 2.5, r_5 = 7.5, r_6 = 6, r_7 = 4, r_8 = 5, r_9 = 2.5, r_{10} = 16.5.$$

   — Statistiques de Wilcoxon et de Mann-Whitney

$$W = \sum_{i=1}^{10} r_i = 64 \text{ et } U = W - \frac{m(m+1)}{2} = 64 - 55 = 9.$$

   — $p-$valeur

   Comme $m = n \geq 8$, on peut utiliser l'approximation normale de la loi de U. Cette approximation conduit à la $p-$valeur suivante

$$\text{p-val} = 2\left[1 - F\left(\frac{|U - E[U]| - 0.5}{\sqrt{\text{var}[U]}}\right)\right] = 2\left[1 - F\left(\frac{|9 - 50| - 0.5}{\sqrt{175}}\right)\right]$$

   car $E[U] = \frac{nm}{2} = 50$ et $\text{var}[U] = \frac{nm(n+m+1)}{12}$. Pour déterminer si les deux échantillons sont significativement différents, on peut calculer la p-valeur et la comparer à un risque de première espèce donné, par exemple $\alpha = 0.05$. Si la $p-$valeur est inférieure à $\alpha = 0.05$, on rejette l'hypothèse H_0 avec ce risque et donc on décide que les deux échantillons sont significativement différents. Si cette $p-$valeur est supérieure à $\alpha = 0.05$, on accepte l'hypothèse H_0 avec ce risque et donc on décide que les deux échantillons ne sont pas significativement différents. Des calculs numériques permettraient d'obtenir (non demandé dans l'examen)

$$\text{p-val} = 2[1 - F(3.0615)] \approx 0.0022.$$

   Comme la $p-$valeur est inférieure à 0.05, on rejette H_0 avec $\alpha = 0.05$.

2. Pour confirmer les résultats obtenus à la question précédente, on regroupe les différentes notes en 4 classes $C_1 = \{0, ..., 4\}$, $C_2 = \{5, ..., 10\}$, $C_3 = \{11, ..., 15\}$ et $C_4 = \{16, ..., 20\}$ pour obtenir le tableau suivant

	Groupe 1	Groupe 2	$N_{k.}$
C_1	0	3	3
C_2	1	5	6
C_3	4	1	5
C_4	5	1	6
$N_{.l}$	10	10	20

Exprimer la valeur de la statistique du test du χ^2 d'homogénéité en fonction des données du problème (sans chercher à la calculer) et sa loi asymptotique. Déterminer le seuil de détection en fonction de l'inverse de la fonction de répartition d'une loi du χ^2_K notée F_K^{-1} pour une valeur de K que l'on déterminera. Expliquer comment déterminer si les résultats des deux groupes sont significativement différents ou pas.
Le test du χ^2 d'homogénéité rejette l'hypothèse H_0 si

$$I_n = \sum_{k=1}^{K}\sum_{l=1}^{L} \frac{\left(N_{k,l} - \frac{N_{k,.}N_{.,l}}{n}\right)^2}{\frac{N_{k,.}N_{.,l}}{n}} > S_{K,L,\alpha}$$

avec $S_{K,L,\alpha} = F^{-1}_{(K-1)(L-1)}(\alpha)$. Dans notre exemple, on a

$$I_n = \frac{\left(0 - \frac{3\times10}{20}\right)^2}{\frac{3\times10}{20}} + \frac{\left(1 - \frac{6\times10}{20}\right)^2}{\frac{6\times10}{20}} + \frac{\left(4 - \frac{5\times10}{20}\right)^2}{\frac{5\times10}{20}} + \frac{\left(5 - \frac{6\times10}{20}\right)^2}{\frac{6\times10}{20}} + \frac{\left(3 - \frac{3\times10}{20}\right)^2}{\frac{3\times10}{20}} + \frac{\left(5 - \frac{6\times10}{20}\right)^2}{\frac{6\times10}{20}} + \frac{\left(1 - \frac{5\times10}{20}\right)^2}{\frac{5\times10}{20}} + \frac{\left(1 - \frac{6\times10}{20}\right)^2}{\frac{6\times10}{20}},$$

c'est-à-dire, après calcul (non demandé dans l'examen) $I_n = \frac{152}{15} \approx 10.13$. De plus, pour $\alpha = 0.05$, on a

$$S_{K,L,\alpha} = F_3^{-1}(0.95).$$

Le principe du test du χ^2 d'homogénéité est de rejeter l'hypothèse H_0 (et donc de décider que les deux échantillons n'ont pas la même loi) si $I_n > S_{K,L,\alpha}$. Puisque $S_{K,L,0.05} \approx 0.35$, on rejette H_0 avec le risque $\alpha = 0.05$. Il y a donc un lien entre les résultats obtenus et le groupe de TD, ce qui confirme le résultat obtenu à la question précédente. On remarquera que les effectifs de chaque classe sont très faibles (dans le cours, on recommande des effectifs ≥ 5).

3. **Test de normalité**. On désire tester s'il est raisonnable de supposer que les deux échantillons de notes associées aux deux groupes sont distribués suivant une loi normale ou pas. Expliquer le principe d'un test de normalité que vous pourriez utiliser pour résoudre ce problème.
Réponse : voir cours.