EXERCICES TESTS NON PARAMÉTRIQUES (AVEC CORRECTIONS)

Janvier 2022

Exercice 1 (de Miodrag Sljukic avec correction)

- **Énoncé** : A company sells the same product in two types of stores: classical and self-service stores. The data about income earned in each type of store are as follows:
  - Classical stores: 50, 50, 60, 70, 75, 80, 90, 85
  - Self-service: 55, 75, 80, 90, 105, 65

  On the level of significance of 95%, is there a difference in income among different types of stores?

- **Correction** : Une manière de répondre à cette question est d'effectuer un test de Mann-Whitney. Si on choisit pour x les données des magasins self-service et pour y les données des magasins classiques, les données ordonnées sont

$$y_1 = 50 = y_2 < x_1 = 55 < y_3 = 60 < x_2 = 65 < y_4 = 70 < y_5 = x_3 = 75 < y_6 = x_4 = 80$$
$$< y_7 = 85 < y_8 = x_5 = 90 < x_6 = 105$$

  Donc la somme des rangs des y_j vaut

$$W = 1.5 + 1.5 + 4 + 6 + 7.5 + 9.5 + 11 + 12.5 = 53.5$$

  et la statistique de Mann-Whitney est $U = W - \frac{8\times 9}{2} = 53.5 - 36 = 17.5$.

- Seuils pour $\alpha = 0.05$: comme on veut savoir si les revenus sont différents (et pas si l'un est supérieur ou inférieur à l'autre), on peut faire un test bilatéral. Ce test rejette l'hypothèse H_0 si $U > S_{1,n,m,\alpha}$ ou si $U < S_{2,n,m,\alpha}$. Le premier seuil s'obtient à partir de $P[U > S_{1,n,m,\alpha}] = 0.025$, soit

$$S_{1,n,m,\alpha} = E[U] + \sigma F^{-1}(0.975)$$

  avec $E[U] = \frac{nm}{2} - 0.5 = 23.5$ (le terme -0.5 vient de la correction de continuité) et $\sigma^2 = \frac{nm(n+m+1)}{12} = 60$, d'où $S_{1,n,m,\alpha} \approx 38.68$. Le second seuil s'obtient de manière similaire

$$S_{1,n,m,\alpha} = E[U] + \sigma F^{-1}(0.025) \approx 8.32.$$

  Donc on accepte H_0 avec $\alpha = 0.05$. Les revenus ne dépendent pas du type de magasins. On remarquera que puisque certaines valeurs des observations sont identiques, le test de Kolmogorov-Smirnov n'est pas adapté.

- $p-$value : en utilisant l'approximation normale de la loi de Mann-Whitney avec correction de continuité, on obtient

$$p\text{-value} = 2F\left(\frac{U_{\text{obs}} - E[U]}{\sigma}\right) = 2F\left(\frac{17.5 - 24 - 0.5}{\sqrt{60}}\right) \approx 0.438.$$

  En utilisant les commandes Matlab
  - x1 = [50 50 60 70 75 80 90 85]
  - x2 = [55 75 80 90 105 65]
  - *[p,h,stats] = ranksum(x1,x2,'tail','both')*

  on obtient une p-valeur proche de 0.428 qui est légèrement différente de celle obtenue avec l'approximation découlant de la loi normale (approximation en théorie valable pour des tailles d'échantillons supérieures à 8, ce qui n'est pas le cas ici pour x_2). La p-valeur est supérieure à 0.05, ce qui confirme qu'on accepte H_0 avec $\alpha = 0.05$. Les revenus ne dépendent pas du type de magasins.

Exercice 2 **(de Miodrag Sljukic avec corrections)**

- **Énoncé** : A company conducted a survey in order to examine if the frequency of usage of company's service depends on the size of the city where it's clients live. The summary of survey is given in the following table:

City Size/Frequency of service usage	Always	Sometimes	Never
Small	151	252	603
Medium	802	603	405
Large	753	55	408

Does the frequency of usage of company's service depend on the size of the city?

- **Correction** : On peut faire un test d'homogénéité pour déterminer si les échantillons associés à différentes tailles de villes, soit $\boldsymbol{X}_1 = (151, 252, 603)$, $\boldsymbol{X}_2 = (802, 603, 405)$ et $\boldsymbol{X}_3 = (753, 55, 408)$ ont la même loi ou pas. On a le tableau de contingences suivant

City Size/Frequency of service usage	Always	Sometimes	Never	$N_{k.}$
Small	151	252	603	1006
Medium	802	603	405	1810
Large	753	55	408	1216
$N_{.l}$	1706	910	1416	$n = 4032$

La statistique de test est donc

$$J_n = \frac{\left(151 - \frac{1006\times1706}{4032}\right)^2}{\frac{1006\times1706}{4031}} + ... + \frac{\left(408 - \frac{1216\times1416}{4032}\right)^2}{\frac{1216\times1416}{4032}} = 821.31.$$

Le seuil de décision est $S_\alpha = F_1^{-1}(0.95) = 9.49$. où F_4 est la fonction de répartition d'une loi du χ_4^2 car $(K-1)\times(L-1) = 4$.
On rejette donc H_0 avec le risque $\alpha = 0.05$ et donc on décide que les fréquences d'utilisation des services dépend de la taille des villes (Matlab indique une p-valeur proche de $p = 0$).

Exercice 3 **(de Miodrag Sljukic avec correction)**

- **Énoncé** : A company produces several models of the same product. A survey which was conducted included 200 buyers who were asked about factor that had the strongest influence on their decision to buy a product. The following data summarizes the survey:

Characteristics/Gender	Male	female
Price	301	502
Design	353	155
color	558	153

On the level of significance of 95%, is there a difference between genders in regard to characteristics of this product?

- **Correction** : On peut faire un test d'indépendance pour déterminer si les goûts des hommes et les femmes sont corrélés ou indépendants. On a le tableau de contingences suivant

Characteristics/Gender	Male	Female	$N_{k.}$
Price	301	502	803
Design	353	155	508
Color	558	153	711
$N_{.l}$	1212	810	$n = 2022$

La statistique de test est donc

$$J_n = \frac{\left(301 - \frac{803\times1212}{2022}\right)^2}{\frac{803\times1212}{2022}} + ... + \frac{\left(153 - \frac{711\times810}{2022}\right)^2}{\frac{711\times810}{2022}} = 289.71.$$

Le seuil de décision est $S_\alpha = F_2^{-1}(0.95) = 5.99$. où F_2 est la fonction de répartition d'une loi du χ_2^2 car $(K-1)\times(L-1) = 2$.
On rejette donc H_0 avec le risque $\alpha = 0.05$ et donc on décide que les hommes et les femmes ont des goûts corrélés (Matlab indique une p-valeur proche de $p = 0$).

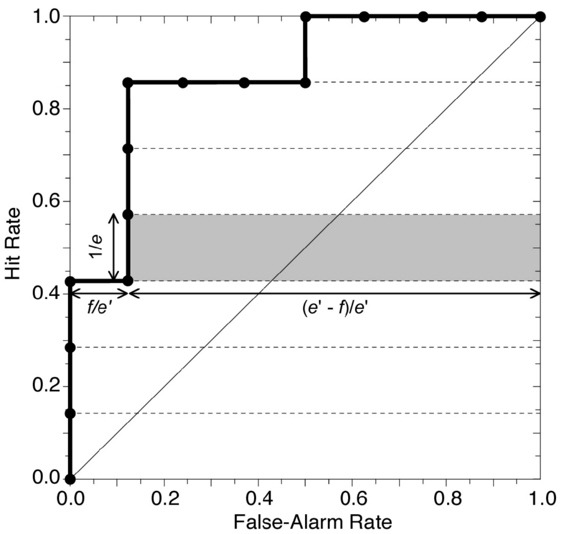
Exercice 4 (de JYT avec correction) : lien entre courbe COR et statistique du test de Mann-Whitney

- **Énoncé** : Cet exercice est inspiré d'un article de S. J. Mason et N. E. Graham cité en référence. Les données du tableau 2 ci-dessous indiquent pour chaque année la probabilité d'avoir une précipitation au nord-est du Brésil notée P à l'aide d'un modèle de prédiction météorologique (colonne de droite) et la réalisation de cet événement ("0" signifie qu'il n'y a pas eu de pluie et "1" signifie qu'il y a eu de pluie). On considère un test statistique qui prédit la présence de pluie si la probabilité π dépasse un seuil S_α qui dépend de la probabilité de fausse alarme $\alpha =$ PFA. Pour une valeur de seuil donnée, on décide d'estimer les probabilités de détection $\pi = 1 - \beta =$ PD et de fausse alarme PFA en comptant le nombre d'évènements dépassant sous chacune des deux hypothèses. En faisant varier le seuil S_α dans l'intervalle $]0, 1[$, on obtient la courbe COR représentée sur la figure ci-dessous.

TABLE 2. FORECASTS FROM TABLE 1 SORTED BY DECREASING ELVIS-ED (SEE APPENDIX) FORECAST PROBABILITY

Year	Event (1)/ non-event (0)	Probability (%)
1994	1	98.4
1995	1	95.2
1984	1	94.4
1981	0	92.8
1985	1	83.2
1986	1	81.6
1988	1	58.4
1982	0	57.6
1991	0	28.0
1987	0	13.6
1989	1	3.2
1992	0	2.4
1990	0	1.6
1983	0	0.8
1993	0	0.0

Estimation de la courbe COR pour les données du tableau 2.



  - Déterminer les coordonnées du point (PFA, PD) associé au seuil $S_\alpha = 0.4$ de la courbe COR.
  - Vérifier que lorsqu'on augmente le seuil S_α de manière à dépasser une donnée associée à l'évènement "0", la probabilité PFA diminue de $\frac{1}{e'}$ et la probabilité PD reste constante, où e' est le nombre d'évènements "0".

- De même, montrer que lorsqu'on augmente le seuil S_α de manière à dépasser une donnée associée à l'évènement "1", la probabilité PFA reste constante et la probabilité PD diminue de $\frac{1}{e}$, où e est le nombre total d'évènements "1".
- En déduire l'aire sous la courbe COR représentée sur la figure.
- Déterminer la statistique de Mann-Whitney U associée à $(X_1, ..., X_7)$ et $(Y_1, ..., Y_8)$ où les X_i correspondent aux probabilités associées aux événements "pluie" et les les Y_j correspondent aux probabilités associées aux événements "non pluie". Vérifier que $A = 1 - \frac{U}{ee'}$ et commenter.

- **Correction** :
  - On a PFA $= P[\pi > S_\alpha = |H_0 \text{ vraie}]$. Donc pour $S_\alpha = 0.4$, le nombre d'évènements "0" vérifiant $\pi > S_\alpha$ est $n_{0.4} = 2$ d'où PFA $\approx \frac{2}{e'}$ où $e' = 8$ est le nombre d'évènements "0" du tableau. On obtient donc PFA ≈ 0.25. De la même façon, on obtient PD $\approx \frac{6}{7} \approx 0.86$. Le point $(0.25, 0.86)$ est bien un point de la courbe COR représentée ci-dessus.
  - Lorsqu'on augmente le seuil S_α de manière à dépasser une donnée associée à l'évènement "0", le nombre de fausses alarmes diminue de 1 donc la probabilité PFA diminue de $\frac{1}{e'}$. La probabilité PD reste constante car le nombre de détections n'a pas changé.
  - Lorsqu'on augmente le seuil S_α de manière à dépasser une donnée associée à l'évènement "1", le nombre de détections diminue de 1 donc la probabilité PD diminue de $\frac{1}{e}$, où e est le nombre total d'évènements "1". la probabilité PFA reste constante car le nombre de fausses alarmes n'a pas changé.
  - L'aire sous la courbe COR est
$$A = \left(3 \times \frac{1}{7}\right) + \left(3 \times \frac{1}{7} \times \left(1 - \frac{1}{8}\right)\right) + \left(\frac{1}{7} \times \frac{4}{8}\right) = \frac{7}{8}.$$
  - La statistique du test de Mann-Whitney s'écrit
$$U = W - \frac{m(m+1)}{2} = W - \frac{m(m+1)}{2} = W - \frac{8 \times 9}{2} = W - 36$$
où W est la somme des rangs des Y_j. Puisque
$$y_1 = 0 < y_2 = 0.8 < y_3 = 1.6 < y_4 = 2.4 < x_1 = 3.2 < ... < x_4 = 83.2 < y_8 = 92.82$$
on a
$$W = 1 + 2 + 3 + 4 + 6 + 7 + 8 + 12 = 43$$
donc $U = 43 - 36 = 7$. On vérifie bien que
$$1 - \frac{U}{ee'} = 1 - \frac{7}{7 \times 8} = \frac{7}{8} = A.$$
Ce résultat est général et illustre le faite que l'aire sous la courbe COR est reliée à la statistique du test de Mann-Whitney. Quand U augmente, les échantillons $(X_1, ..., X_7)$ et $(Y_1, ..., Y_8)$ de plus en plus différents et donc l'aire sous la courbe COR augmente, ce qui signifie qu'il sera plus facile de détecter la présence de pluie.

Exercice 5 (inspiré du sujet d'examen d'Anne Fromont de 2011-2012 avec correction)

- **Énoncé** : La loi de Benford découverte en 1938 stipule que la probabilité que le premier chiffre C d'un nombre issu de résultats de mesures est définie comme dans le tableau suivant

k	1	2	3	4	5	6	7	8	9
P[C=k]	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Une personne relève 120 montants de factures reportés dans le tableau ci-dessous et se demande si ces montants ont été falsifiés ou pas. Pour répondre à cette question, effectuer un test d'adéquation du χ^2 avec le risque $\alpha = 0.05$. Déterminer la p-valeur de ce test. Qu'en concluez vous ?

k	1	2	3	4	5	6	7	8	9
Montants commençant par k	12	10	12	13	17	18	16	12	10

- **Correction** : Le test du χ^2 d'adéquation est défini par

$$\text{Rejet de } H_0 \text{ si } \phi_n = \sum_{k=1}^{K} \frac{(Z_k - np_k)^2}{np_k} > S_{K,\alpha}$$

où l'hypothèse H_0 est "les factures n'ont pas été falsifiées et sont en accord avec la loi de Benford. La statistique de test est donc

$$\phi_n = \frac{(12 - 120 \times 0.301)^2}{120 \times 0.301} + \frac{(10 - 120 \times 0.176)^2}{120 \times 0.176} + ... + \frac{(10 - 120 \times 0.046)^2}{120 \times 0.046} = \frac{608}{45} \approx 62.05.$$

Sous l'hypothèse H_0, ϕ_n suit asymptotiquement une loi du χ_8^2, donc

$$S_{K,\alpha} = F_8^{-1}(1 - \alpha)$$

où F_8^{-1} est l'inverse de la fonction de répartition d'une loi du χ_8^2. Pour $\alpha = 0.05$, on obtient

$$S_{K,\alpha} = F_3^{-1}(0.95) \approx 15.507.$$

On rejette donc l'hypothèse H_0 avec le risque $\alpha = 0.05$ et on en conclut que les si la loi de Benford est correcte, les factures ont été falsifiées. On remarquera que le premier écart du test du χ^2 est

$$\frac{(12 - 120 \times 0.301)^2}{120 \times 0.301} = 16.11 > S_{K,\alpha} = 15.507$$

ce qui montre que la décision est franche. La p-valeur du test est

$$p = 1 - F_8(\phi_n) \approx 1.84e{-10}$$

qui confirme que la décision est franche.

Exercice 6 **(inspiré du sujet d'examen d'Anne Fromont de 2009-2010 avec correction)**

- **Énoncé** : On se demande ici si le choix parental de vacciner ses enfants est lié ou non à l'âge des enfants. Pour cela, on relève pour 200 enfants choisis au hasard dans la population, la tranche d'âge des enfants ainsi que le nombre d'enfants vaccinés par tranche d'âge. On obtient les résultats suivants :

Vaccination/Tranche âge	moins de 9 ans	plus de 9 ans
Vaccinés	27	7
Non vaccinés	99	67

  À l'aide d'un test du χ^2 d'indépendance de niveau asymptotique 5% que l'on décrira et justifiera avec précision, déterminer si le choix parental de vaccination des enfants est lié ou non à l'âge des enfants.

- **Correction** :

On peut faire un test d'indépendance pour déterminer si le choix parental de vaccination des enfants est lié ou non à l'âge des enfants. On a le tableau de contingences suivant

Vaccination/Tranche âge	moins de 9 ans	plus de 9 ans	$N_{k.}$
Vaccinés	27	7	34
Non vaccinés	99	67	166
$N_{.l}$	126	74	$n = 200$

La statistique de test est donc

$$J_n = \frac{\left(27 - \frac{34\times126}{200}\right)^2}{\frac{34\times126}{200}} + ... + \frac{\left(67 - \frac{166\times74}{200}\right)^2}{\frac{166\times74}{200}} \approx 4.73.$$

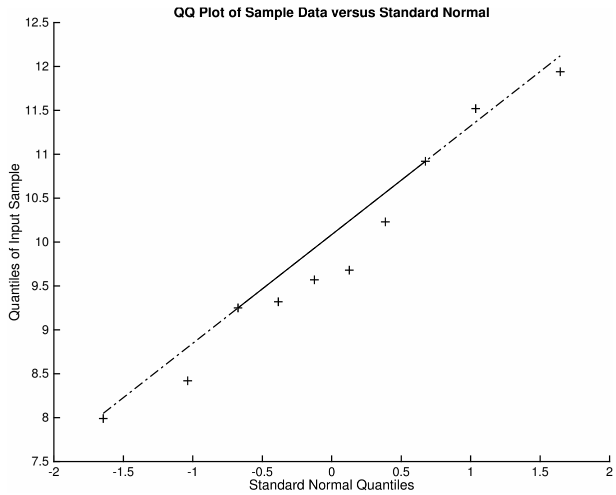
Le seuil de décision est $S_\alpha = F_1^{-1}(0.95) \approx 3.84.$ où F_1 est la fonction de répartition d'une loi du χ_1^2 car $(K-1)\times(L-1) = 1$.
On rejette donc H_0 avec le risque $\alpha = 0.05$ et donc on décide que le choix parental de vaccination des enfants est lié à l'âge des enfants avec un risque $\alpha = 0.05$ (Matlab indique une p-valeur $p = 0.0296$).

Exercice 7 (de JYT) : tests de normalité avec correction

- **Énoncé** : L'objectif de cet exercice est de comparer les résultats de tests de normalité pour l'échantillon suivant

8.42	7.99	10.23	10.92	11.94	9.32	11.52	9.25	9.57	9.68

  - Un élève a utilisé la commande qqplot(x) sous Matlab et obtient le résultat de la figure ci-dessous. Pouvez vous expliquer ce que représente ce graphique. Peut-on conclure quant-à la normalité de l'échantillon ?



  - Que donne le test de Lilliefors pour cet échantillon ?
  - Expliquer le principe du test de Shapiro-Wilk et calculer la statistique de ce test en utilisant les valeurs de $a_{[i]}$ de la table ci-dessous

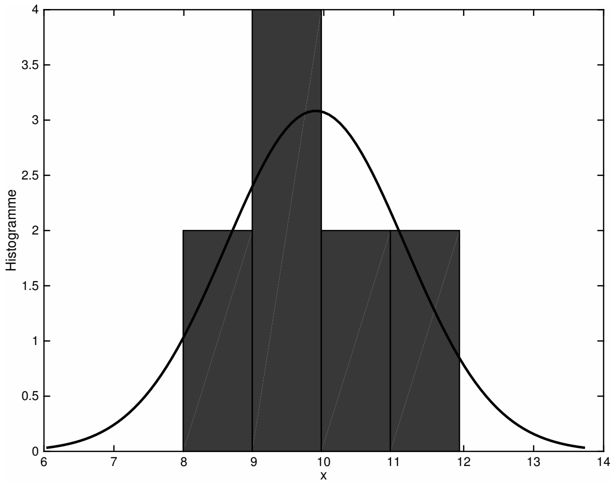
Table 7 : Coefficients de Shapiro-Wilk :
n = taille de l'échantillon, i = numéro de la différence d_i

i / n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	0,7071														
3	0,7071	0													
4	0,6872	0,1677													
5	0,6646	0,2413	0												
6	0,6431	0,2806	0,0875												
7	0,6233	0,3031	0,1401	0											
8	0,6052	0,3164	0,1743	0,0561											
9	0,5888	0,3244	0,1976	0,0947	0										
10	0,5739	0,3291	0,2141	0,1224	0,0399										
11	0,5601	0,3315	0,226	0,1429	0,0695	0									
12	0,5475	0,3325	0,2347	0,1586	0,0922	0,0303									
13	0,5359	0,3325	0,2412	0,1707	0,1099	0,0539	0								
14	0,5251	0,3318	0,246	0,1802	0,124	0,0727	0,024								
15	0,515	0,3306	0,2495	0,1878	0,1353	0,088	0,0433	0							
16	0,5056	0,329	0,2521	0,1939	0,1447	0,1005	0,0593	0,0196							
17	0,4963	0,3273	0,254	0,1988	0,1524	0,1109	0,0725	0,0359	0						
18	0,4886	0,3253	0,2553	0,2027	0,1587	0,1197	0,0837	0,0496	0,0163						
19	0,4808	0,3232	0,2561	0,2059	0,1641	0,1271	0,0932	0,0612	0,0303	0					
20	0,4734	0,3211	0,2565	0,2085	0,1686	0,1334	0,1013	0,0711	0,0422	0,014					
21	0,4643	0,3185	0,2578	0,2119	0,1736	0,1399	0,1092	0,0804	0,053	0,0263	0				
22	0,459	0,3156	0,2571	0,2131	0,1764	0,1443	0,115	0,0878	0,0618	0,0368	0,0122				
23	0,4542	0,3126	0,2563	0,2139	0,1787	0,148	0,1201	0,0941	0,0696	0,0459	0,0228	0			
24	0,4493	0,3098	0,2554	0,2145	0,1807	0,1512	0,1245	0,0997	0,0764	0,0539	0,0321	0,0107			
25	0,445	0,3069	0,2543	0,2148	0,1822	0,1539	0,1283	0,1046	0,0823	0,061	0,0403	0,02	0		
26	0,4407	0,3043	0,2533	0,2151	0,1836	0,1563	0,1316	0,1089	0,0876	0,0672	0,0476	0,0284	0,0094		
27	0,4366	0,3018	0,2522	0,2152	0,1848	0,1584	0,1346	0,1128	0,0923	0,0728	0,054	0,0358	0,0178	0	
28	0,4328	0,2992	0,251	0,2151	0,1857	0,1601	0,1372	0,1162	0,0965	0,0778	0,0598	0,0424	0,0253	0,0084	
29	0,4291	0,2968	0,2499	0,215	0,1064	0,1616	0,1395	0,1192	0,1002	0,0822	0,065	0,0483	0,032	0,0159	0
30	0,4254	0,2944	0,2487	0,2148	0,187	0,163	0,1415	0,1219	0,1036	0,0862	0,0697	0,0537	0,0381	0,0227	0,0076

Table des valeurs limites de W

n	5%	1%
3	0.767	0.753
4	0.748	0.687
5	0.762	0.686
6	0.788	0.713
7	0.803	0.730
8	0.818	0.749
9	0.829	0.764
10	0.842	0.781
11	0.850	0.792
12	0.859	0.805
13	0.856	0.814
14	0.874	0.825
15	0.881	0.835
16	0.837	0.844
17	0.892	0.851
18	0.897	0.858
19	0.901	0.863
20	0.905	0.868
21	0.908	0.873
22	0.911	0.878
23	0.914	0.881
24	0.916	0.884
25	0.918	0.888
26	0.920	0.891
27	0.923	0.894
28	0.924	0.896
29	0.926	0.898
30	0.927	0.900

- **Correction** : avec un nombre de données aussi faible, la superposition de l'histogramme et de la densité de probabilité de la loi normale de moyenne $\hat{m} = \bar{x}$ et de variance $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ est peu informative, comme le montre la figure ci-dessous



Il est donc plus judicieux d'appliquer un des tests de normalité du cours.

  - Q-Q plot : le Q-Q plot représente l'ensemble des points $M_i = (x_{(i)}, t_{(i)})$ où $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ est la statistique d'ordre de l'échantillon $(x_1, ..., x_n)$ et $t_{(i)} = \Phi^{-1}\left(\frac{i}{n}\right)$, Φ étant la fonction de répartition de la loi normale $\mathcal{N}(0,1)$. Dans le cas d'un échantillon de loi normale, les points M_i doivent être alignés. Dans notre exemple, il est difficile de conclure.
  - Test de Lilliefors : la valeur de la statistique de test est $D_n = 0.1634$ et le seuil associé est $S_{10,0.05} \approx 0.26$ donc on accepte l'hypothèse de normalité avec le risque $\alpha = 0.05$ (la p-valeur vérifie $p = 0.50$). Notons que le seuil peut être déterminé numériquement ou à l'aide des tables suivantes

    Table de Lilliefors

n / α	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	30	n>30
0,1	0,352	0,15	0,294	0,276	0,261	0,249	0,239	0,23	0,223	0,214	0,207	0,201	0,195	0,189	0,184	0,179	0,174	0,158	0,144	0,805/$\sqrt{n}$
0,05	0,381	0,337	0,19	0,3	0,285	0,271	0,258	0,249	0,243	0,234	0,227	0,22	0,213	0,206	0,2	0,195	0,19	0,173	0,161	0,886/$\sqrt{n}$
0,01	0,417	0,405	0,364	0,348	0,311	0,311	0,294	0,284	0,242	0,268	0,261	0,257	0,25	0,245	0,239	0,235	0,231	0,2	0,187	1,031/$\sqrt{n}$

    La commande Matlab est $x = [247.0, 247.8, 250.2, 251.3, 251.9, 249.4, 248.8, 247.1, 255.0, 247.0, 254.8, ...]$ et *[H,P,KSTAT,critval] = lillietest(x)*.
  - Test de Shapiro-Wilk : la statistique du test de Shapiro-Wilk est définie par

$$\mathrm{SW}_n = \frac{\left[\sum_{i=1}^{\mathrm{ent}(\frac{n}{2})} a_i (X_{(n-i+1)} - X_{(i)})\right]^2}{\sum_{i=1}^{n} (X_i - \bar{X}_n)^2}.$$

    Le numérateur de cette statistique est égal à

$$a_1[X_{(10)} - X_{(1)}]^2 + a_2[X_{(9)} - X_{(2)}]^2 + a_3[X_{(8)} - X_{(3)}]^2 + a_4[X_{(7)} - X_{(4)}]^2 + a_5[X_{(6)} - X_{(5)}]^2 \approx 14.1409$$

    tandis que le dénominateur vaut

$$\sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \approx 14.6874$$

    d'où $\mathrm{SW}_n \approx 0.963$. Le seuil associé est $S_{10,0.05} = 0.842$ donc on accepte l'hypothèse de normalité avec le risque $\alpha = 0.05$. Si on utilise un logiciel comme Matlab, on obtient le même résultat avec une p-valeur égale à $p = 0.8152$).

    La commande Matlab est *[H, pValue, SWstatistic] = swtest(x, 0.05)*.

Exercice 8 **(de JYT avec correction) : comparaison de deux détecteurs d'anomalies**

- **Énoncé** : On souhaite comparer la performance de deux détecteurs d'anomalies qui permettent pour chaque vecteur de données de calculer un score d'anomalie. Les résultats obtenus pour ces deux détecteurs sur les mêmes jeux de données (rangés par ordre croissant) sont reportés dans le tableau ci-dessous

Scores détecteur 1	0.00	0.03	0.05	0.09	0.16	0.19	0.20	0.27	0.30	0.33
	0.36	0.38	0.62	0.74	0.76	0.78	0.79	0.85	0.88	0.94
Scores détecteur 2	0.03	0.05	0.13	0.17	0.18	0.27	0.30	0.42	0.43	0.54
	0.55	0.56	0.66	0.67	0.69	0.70	0.74	0.94	0.98	0.99

  Il semblerait que le second détecteur fournit des scores d'anomalies plus élevés mais on aimerait le justifier à l'aide de tests statistiques.

  – Le test de Student semble adapté pour résoudre ce problème mais il nécessite que les deux échantillons soient gaussiens. Effectuer le test de Shapiro-Wilk pour ces deux jeux de données et conclure.

  – Au vu des résultats de la question précédente, on s'oriente vers l'utilisation de tests non-paramétriques. Utiliser le test de Mann-Whitney pour déterminer si les scores d'anomalie du second détecteur sont supérieurs à ceux du premier détecteur. On déterminera la valeur de la statistique de test, le seuil pour un risque $\alpha = 0.05$, la $p-$valeur de ce test et on prendra une décision pour cette valeur du risque α.

- **Correction** :
  Le test de **Shapiro-Wilk** appliqué aux deux jeux de données notés $(x_1, ..., x_n)$ et $(y_1, ...y_n)$ avec $n = 20$ fournit les résultats suivants

  – Détecteur 1 : p-value = 0.0372, SWstatistic = 0.8976, $S_{n,\alpha} = 0.905$ avec $\alpha = 0.05$.

  – Détecteur 2 : p-value = 0.3544, SWstatistic = 0.9492, $S_{n,\alpha} = 0.905$ avec $\alpha = 0.05$.

  On peut donc conclure que le le détecteur 1 fournit des scores d'anomalies qui ne peuvent être considérés comme gaussiens avec le risque $\alpha = 0.05$ tandis que l'hypothèse de normalité peut être acceptée pour le détecteur 2 avec le même risque $\alpha = 0.05$. On ne peut donc pas utiliser le test de Student.

  Le test de **Mann-Whitney** consiste à tester si la loi des scores du second détecteur de fonction de répartition G vérifie $G > F$, où F est la fonction de répartition des scores du premier détecteur. Le détecteur rejette l'hypothèse H_0 (les scores des deux détecteurs ont la même loi) si

$$U = W - \frac{m(m+1)}{2} > S_{n,m,\alpha}$$

  où $W = \sum_{j=1}^{m} S_j$ et S_j est le rang de Y_j parmi les $n + m$ données réunies $(X_1, ..., X_n, Y_1, ..., Y_m)$ (le minimum a le rang 1, la suivante le rang 2, ...). On obtient les résultats suivants

  – Somme des rangs et statistique de Mann-Whitney : $W = 429$ et $U_{\text{obs}} = W - \frac{20\times 21}{2} = 219$.

  – Seuil pour $\alpha = 0.05$:

$$S_{\alpha,n,m} = E[U] + \sigma F^{-1}(0.95)$$

  avec $E[U] = \frac{nm}{2} = 200$ et $\sigma^2 = \frac{nm(n+m+1)}{12} = \frac{4100}{3}$, d'où $S_{0.05,20,20} \approx 260.8$.

  – $p-$value : en utilisant l'approximation normale de la loi de Mann-Whitney avec correction de continuité, on obtient

$$p\text{-value} = 1 - F\left(\frac{U_{\text{obs}} - E[U]}{\sigma}\right)$$

  avec $E[U] = \frac{nm}{2} - 0.5$, $\sigma^2 = \frac{nm(n+m+1)}{12}$ et F est la fonction de répartition de la loi normale $\mathcal{N}(0, 1)$. Après application numérique, on obtient p-value ≈ 0.30.

  – Commande sous Matlab : [p,h,stats] = ranksum(d2,d1,'tail','right') car sous Matlab, la routine ranksum calcule les rangs de la première variable.

  Donc on accepte l'hypothèse H_0 avec $\alpha = 0.05$ et on décide que les deux détecteurs ont des scores d'anomalies similaires.

Références

- Exercices de Miodrag Sljukic : https://www.r-exercises.com/2016/11/20/nonparametric-tests/
- S. J. Mason and N. E. Graham, Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation, Quaterly Journal Royal of the Meteorological Society, vol. 128, no. 584, pp. 2145-2166, July 2002.
- Page personnelle d'Alessio Guarino : https://blog.univ-reunion.fr/alessioguarino/2017/11/16/mann-whitney-wilcoxon-test-examples/
- Polycopié de cours de Magali Fromont, "Tests statistiques : rejeter, ne pas rejeter ... Se risquer ?", Université de Rennes 2, Année universitaire 2015-2016.

$n \backslash \alpha$	0.01	0.05	0.10	0.15	0.20
16	0.2477	0.2128	0.1956	0.1843	0.1758
17	0.2408	0.2071	0.1902	0.1794	0.1711
18	0.2345	0.2018	0.1852	0.1747	0.1666
19	0.2285	0.1965	0.1803	0.1700	0.1624
20	0.2226	0.1920	0.1764	0.1666	0.1589
21	0.2190	0.1881	0.1726	0.1629	0.1553
22	0.2141	0.1840	0.1690	0.1592	0.1517
23	0.2090	0.1798	0.1650	0.1555	0.1484
24	0.2053	0.1766	0.1619	0.1527	0.1458
25	0.2010	0.1726	0.1589	0.1498	0.1429
26	0.1985	0.1699	0.1562	0.1472	0.1406
27	0.1941	0.1665	0.1533	0.1448	0.1381
28	0.1911	0.1641	0.1509	0.1423	0.1358
29	0.1886	0.1614	0.1483	0.1398	0.1334
30	0.1848	0.1590	0.1460	0.1378	0.1315
31	0.1820	0.1559	0.1432	0.1353	0.1291
32	0.1798	0.1542	0.1415	0.1336	0.1274
33	0.1770	0.1518	0.1392	0.1314	0.1254
34	0.1747	0.1497	0.1373	0.1295	0.1236
35	0.1720	0.1478	0.1356	0.1278	0.1220
36	0.1695	0.1454	0.1336	0.1260	0.1203
37	0.1677	0.1436	0.1320	0.1245	0.1188
38	0.1653	0.1421	0.1303	0.1230	0.1174
39	0.1634	0.1402	0.1288	0.1214	0.1159
40	0.1616	0.1386	0.1275	0.1204	0.1147
41	0.1599	0.1373	0.1258	0.1186	0.1131
42	0.1573	0.1353	0.1244	0.1172	0.1119
43	0.1556	0.1339	0.1228	0.1159	0.1106
44	0.1542	0.1322	0.1216	0.1148	0.1095
45	0.1525	0.1309	0.1204	0.1134	0.1083
46	0.1512	0.1293	0.1189	0.1123	0.1071
47	0.1499	0.1282	0.1180	0.1113	0.1062
48	0.1476	0.1269	0.1165	0.1098	0.1047
49	0.1463	0.1256	0.1153	0.1089	0.1040
50	0.1457	0.1246	0.1142	0.1079	0.1030
OVER 50	$\frac{1.035}{f(n)}$	$\frac{0.895}{f(n)}$	$\frac{0.819}{f(n)}$	$\frac{0.775}{f(n)}$	$\frac{0.741}{f(n)}$

where

$$f(n) = \frac{.83 + n}{\sqrt{n}} - .01$$

Valeurs du seuil pour le test de Lilliefors.

Table 4b : table des valeurs limites W_α de $W = \frac{b^2}{Z^2}$

pour les risques α = 5 % et 1 %
(Biometrika 1965)

n	Risque 5 %	Risque 1 %
	$W_{0,05}$	$W_{0,01}$
5	0,762	0,686
6	0,788	0,713
7	0,803	0,730
8	0,818	0,749
9	0,829	0,764
10	0,842	0,781
11	0,850	0,792
12	0,859	0,805
13	0,866	0,814
14	0,874	0,825
15	0,881	0,835
16	0,887	0,844
17	0,892	0,851
18	0,897	0,858
19	0,901	0,863
20	0,905	0,868
21	0,908	0,873
22	0,911	0,878
23	0,914	0,881
24	0,916	0,884
25	0,918	0,888
26	0,920	0,891
27	0,923	0,894
28	0,924	0,896
29	0,926	0,898
30	0,927	0,900
31	0,929	0,902
32	0,930	0,904
33	0,931	0,906
34	0,933	0,908
35	0,934	0,910
36	0,935	0,912
37	0,936	0,914
38	0,938	0,916
39	0,939	0,917
40	0,940	0,919
41	0,941	0,920
42	0,942	0,922
43	0,943	0,923
44	0,944	0,924
45	0,945	0,926
46	0,945	0,927
47	0,946	0,928
48	0,947	0,929
49	0,947	0,929
50	0,947	0,930

Valeurs du seuil pour le test de Shapiro-Wilk.

Loi de Khi-deux

Le tableau donne **x** tel que **P(K > x) = p**

p	0,999	0,995	0,99	0,98	0,95	0,9	0,8	0,2	0,1	0,05	0,02	0,01	0,005	0,001
ddl														
1	0,0000	0,0000	0,0002	0,0006	0,0039	0,0158	0,0642	1,6424	2,7055	3,8415	5,4119	6,6349	7,8794	10,8276
2	0,0020	0,0100	0,0201	0,0404	0,1026	0,2107	0,4463	3,2189	4,6052	5,9915	7,8240	9,2103	10,5966	13,8155
3	0,0243	0,0717	0,1148	0,1848	0,3518	0,5844	1,0052	4,6416	6,2514	7,8147	9,8374	11,3449	12,8382	16,2662
4	0,0908	0,2070	0,2971	0,4294	0,7107	1,0636	1,6488	5,9886	7,7794	9,4877	11,6678	13,2767	14,8603	18,4668
5	0,2102	0,4117	0,5543	0,7519	1,1455	1,6103	2,3425	7,2893	9,2364	11,0705	13,3882	15,0863	16,7496	20,5150
6	0,3811	0,6757	0,8721	1,1344	1,6354	2,2041	3,0701	8,5581	10,6446	12,5916	15,0332	16,8119	18,5476	22,4577
7	0,5985	0,9893	1,2390	1,5643	2,1673	2,8331	3,8223	9,8032	12,0170	14,0671	16,6224	18,4753	20,2777	24,3219
8	0,8571	1,3444	1,6465	2,0325	2,7326	3,4895	4,5936	11,0301	13,3616	15,5073	18,1682	20,0902	21,9550	26,1245
9	1,1519	1,7349	2,0879	2,5324	3,3251	4,1682	5,3801	12,2421	14,6837	16,9190	19,6790	21,6660	23,5894	27,8772
10	1,4787	2,1559	2,5582	3,0591	3,9403	4,8652	6,1791	13,4420	15,9872	18,3070	21,1608	23,2093	25,1882	29,5883
11	1,8339	2,6032	3,0535	3,6087	4,5748	5,5778	6,9887	14,6314	17,2750	19,6751	22,6179	24,7250	26,7568	31,2641
12	2,2142	3,0738	3,5706	4,1783	5,2260	6,3038	7,8073	15,8120	18,5493	21,0261	24,0540	26,2170	28,2995	32,9095
13	2,6172	3,5650	4,1069	4,7654	5,8919	7,0415	8,6339	16,9848	19,8119	22,3620	25,4715	27,6882	29,8195	34,5282
14	3,0407	4,0747	4,6604	5,3682	6,5706	7,7895	9,4673	18,1508	21,0641	23,6848	26,8728	29,1412	31,3193	36,1233
15	3,4827	4,6009	5,2293	5,9849	7,2609	8,5468	10,3070	19,3107	22,3071	24,9958	28,2595	30,5779	32,8013	37,6973
16	3,9416	5,1422	5,8122	6,6142	7,9616	9,3122	11,1521	20,4651	23,5418	26,2962	29,6332	31,9999	34,2672	39,2524
17	4,4161	5,6972	6,4078	7,2550	8,6718	10,0852	12,0023	21,6146	24,7690	27,5871	30,9950	33,4087	35,7185	40,7902
18	4,9048	6,2648	7,0149	7,9062	9,3905	10,8649	12,8570	22,7595	25,9894	28,8693	32,3462	34,8053	37,1565	42,3124
19	5,4068	6,8440	7,6327	8,5670	10,1170	11,6509	13,7158	23,9004	27,2036	30,1435	33,6874	36,1909	38,5823	43,8202
20	5,9210	7,4338	8,2604	9,2367	10,8508	12,4426	14,5784	25,0375	28,4120	31,4104	35,0196	37,5662	39,9968	45,3147
21	6,4467	8,0337	8,8972	9,9146	11,5913	13,2396	15,4446	26,1711	29,6151	32,6706	36,3434	38,9322	41,4011	46,7970
22	6,9830	8,6427	9,5425	10,6000	12,3380	14,0415	16,3140	27,3015	30,8133	33,9244	37,6595	40,2894	42,7957	48,2679
23	7,5292	9,2604	10,1957	11,2926	13,0905	14,8480	17,1865	28,4288	32,0069	35,1725	38,9683	41,6384	44,1813	49,7282
24	8,0849	9,8862	10,8564	11,9918	13,8484	15,6587	18,0618	29,5533	33,1962	36,4150	40,2704	42,9798	45,5585	51,1786
25	8,6493	10,5197	11,5240	12,6973	14,6114	16,4734	18,9398	30,6752	34,3816	37,6525	41,5661	44,3141	46,9279	52,6197
26	9,2221	11,1602	12,1981	13,4086	15,3792	17,2919	19,8202	31,7946	35,5632	38,8851	42,8558	45,6417	48,2899	54,0520
27	9,8028	11,8076	12,8785	14,1254	16,1514	18,1139	20,7030	32,9117	36,7412	40,1133	44,1400	46,9629	49,6449	55,4760
28	10,3909	12,4613	13,5647	14,8475	16,9279	18,9392	21,5880	34,0266	37,9159	41,3371	45,4188	48,2782	50,9934	56,8923
29	10,9861	13,1211	14,2565	15,5745	17,7084	19,7677	22,4751	35,1394	39,0875	42,5570	46,6927	49,5879	52,3356	58,3012
30	11,5880	13,7867	14,9535	16,3062	18,4927	20,5992	23,3641	36,2502	40,2560	43,7730	47,9618	50,8922	53,6720	59,7031
40	17,9164	20,7065	22,1643	23,8376	26,5093	29,0505	32,3450	47,2685	51,8051	55,7585	60,4361	63,6907	66,7660	73,4020
50	24,6739	27,9907	29,7067	31,6639	34,7643	37,6886	41,4492	58,1638	63,1671	67,5048	72,6133	76,1539	79,4900	86,6608
60	31,7383	35,5345	37,4849	39,6994	43,1880	46,4589	50,6406	68,9721	74,3970	79,0819	84,5799	88,3794	91,9517	99,6072
70	39,0364	43,2752	45,4417	47,8934	51,7393	55,3289	59,8978	79,7146	85,5270	90,5312	96,3875	100,4252	104,2149	112,3169
80	46,5199	51,1719	53,5401	56,2128	60,3915	64,2778	69,2069	90,4053	96,5782	101,8795	108,0693	112,3288	116,3211	124,8392
90	54,1552	59,1963	61,7541	64,6347	69,1260	73,2911	78,5584	101,0537	107,5650	113,1453	119,6485	124,1163	128,2989	137,2084
100	61,9179	67,3276	70,0649	73,1422	77,9295	82,3581	87,9453	111,6667	118,4980	124,3421	131,1417	135,8067	140,1695	149,4493
120	77,7551	83,8516	86,9233	90,3667	95,7046	100,6236	106,8056	132,8063	140,2326	146,5674	153,9182	158,9502	163,6482	173,6174
140	93,9256	100,6548	104,0344	107,8149	113,6593	119,0293	125,7581	153,8537	161,8270	168,6130	176,4709	181,8403	186,8468	197,4508
160	110,3603	117,6793	121,3456	125,4400	131,7561	137,5457	144,7834	174,8283	183,3106	190,5165	198,8464	204,5301	209,8239	221,0190
180	127,0111	134,8844	138,8204	143,2096	149,9688	156,1526	163,8682	195,7434	204,7037	212,3039	221,0772	227,0561	232,6198	244,3705
200	143,8428	152,2410	156,4320	161,1003	168,2786	174,8353	183,0028	216,6088	226,0210	233,9943	243,1869	249,4451	255,2642	267,5405
250	186,5541	196,1606	200,9386	206,2490	214,3916	221,8059	231,0128	268,5986	279,0504	287,8815	298,0388	304,9396	311,3462	324,8324
300	229,9634	240,6634	245,9725	251,8637	260,8781	269,0679	279,2143	320,3971	331,7885	341,3951	352,4246	359,9064	366,8444	381,4252
400	318,2596	330,9028	337,1553	344,0781	354,6410	364,2074	376,0218	423,5895	436,6490	447,6325	460,2108	468,7245	476,6064	493,1318
500	407,9470	422,3034	429,3875	437,2194	449,1468	459,9261	473,2099	526,4014	540,9303	553,1268	567,0698	576,4928	585,2066	603,4460
600	498,6229	514,5289	522,3651	531,0191	544,1801	556,0560	570,6680	628,9433	644,8004	658,0936	673,2703	683,5156	692,9816	712,7712
700	590,0480	607,3795	615,9075	625,3175	639,6130	652,4973	668,3308	731,2805	748,3591	762,6607	778,9721	789,9735	800,1314	821,3468
800	682,0665	700,7250	709,8969	720,0107	735,3623	749,1852	766,1555	833,4557	851,6712	866,9114	884,2789	895,9843	906,7862	929,3289
900	774,5698	794,4750	804,2517	815,0267	831,3702	846,0746	864,1125	935,4987	954,7819	970,9036	989,2631	1001,6296	1013,0364	1036,8260

Table du χ^2.