Master 1 Mathématiques Appliquées et Statistique **2021-2022**
parcours MSS et CMI

TP1 - Statistique Non Paramétrique

Tests de Wilcoxon / Mann et Whitney et de Kruskal-Wallis

Exercice 1 **Tests statistiques de Wilcoxon / Mann et Whitney**

Soient les échantillons indépendants

$X = (4.5, 6.6, 7, 6.7, 3.9, 5.7, 5.2)$,

$Y = (4.2, 5.3, 6.4, 5.1, 4.8)$,

Dans un premier temps, on va tester l'homogénéité (*test bilatéral*) des échantillons X et Y avec le test exact à la main, puis comparer le résultat avec celle de la fonction *wilcox.test* dans R.

(a) Calculer à la main l'observation de la statistique W des rangs du test de Wilcoxon.

```
> X <- c(4.5, 6.6, 7, 6.7, 3.9, 5.7,5.2)
> Y <- c(4.2, 5.3, 6.4, 5.1, 4.8)
> # regrouper les deux échantillons en un vecteur long
> all <- c(X,Y)
> all
 [1] 4.5 6.6 7.0 6.7 3.9 5.7 5.2 4.2 5.3 6.4 5.1 4.8
> # chercher les rangs du petit échantillons, qui est Y
> rank(all)
 [1]  3 10 12 11  1  8  6  2  7  9  5  4
> # ramasser ceux à Y
> n1 <- length(X)
> n2 <- length(Y)
> rank(all)[(n1+1):(n1+n2)]
[1] 2 7 9 5 4
> # calculer la somme
> W <- sum(rank(all)[(n1+1):(n1+n2)])
> W
[1] 27
```

(b) Conclure le test pour alpha=0,05, à l'aide d'une table statistique.

Pour n2=5 et n2=7, et un risque de alpha=0,05, on rejette H0 pour toute valeur de W hors l'intervalle [20,45]. Ici, W=27 donc on ne rejette pas H0.

(c) Comparer avec le résultat de la fonction wilcox.test de R.

Histoire : Le test de Wilcoxon / Mann et Whitney a été proposé par Frank Wilcoxon en 1945 et détaillé par Henry Mann et Donald Ransom Whitney en 1947. Dans leurs recherches, différentes statistiques de test s'emploient mais elles se basent toutes sur les rangs.

Soient les échantillons $(X_1, ..., X_{n_1})$ et $(Y_1, ..., Y_{n_2})$, avec $n_2 \leq n_1$:

- La statistique W de Wilcoxon : $W = \sum_{i2=1}^{n_2} R_{i2}$

- La statistique U de Mann et Whitney : $U = \sum_{i=1}^{n_2} \#\{X_j : X_j < Y_i, j = 1, ..., n_1\}$

Leur relation : $U = W - n_2(n_2 + 1)/2$.

Celle utilisée dans R est U. Etant donée l'observation de la W précedante, on a donc l'observation de la statistique U

```
> U = W - n2*(n2+1)/2
> U
```

```
[1] 12
```

Complèter les argument dans la fonction wilcox.test ci-dessous à l'aide de la page

https ://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/wilcox.test

puis effectuer le test exact de wilcoxon / mann et whitney.

wilcox.test(Y,X,alternative = ,exact=)

NB : on doit placer le petit échantillon en premier, ici Y.

```
> wilcox.test(Y,X,alternative = c("two.sided"),exact=TRUE)
```

```
        Wilcoxon rank sum test

data:  Y and X
W = 12, p-value = 0.4318
alternative hypothesis: true location shift is not equal to 0
```

La valeur de W ainsi que la p-valeur conduitent tous à ne pas rejetter H0.

(d) On va maintenant tester l'homogénéité (*test bilatéral*) de X et Y avec le test asymptotique.

Générer un échantillon de taille $n_1 = 80$ de variables aléatoires i.i.d. selon une loi continue de votre choix. En générer un second de taille $n_2 = 50$ selon une autre loi continue. Tester l'homogénéité de ces deux échantillons à l'aide du test de Mann-Whitney-Wilcoxon.

```
> #théoriquement
> n1 = 80
> n2 = 50
> X <- rnorm(n1, 0)
> Y <- rnorm(n2, 1)
```

(e) calculer W à la main, son espérance EW et sa variance VW sous H0.

```
> # regrouper les deux échantillons en un vect long
> all <- c(X,Y)
> # calculer la somme
> W <- sum(rank(all)[(n1+1):(n1+n2)])
> W
[1] 4297
> #théoriquement
> EW <- n2*(n1+n2+1)/2
> EW
[1] 3275
> VW <- n1*n2*(n1+n2+1)/12
> VW
[1] 43666.67
```

(f) Calculer à la main les valeurs critiques pour W à $\alpha = 0,05$ (rappeler la loi asymptotique de W sous H0) et conclure le test.
$W \overset{n\to\infty}{\sim} \mathcal{N}(EW, VW)$ sous H0.

```
> c1 <- qnorm(0.025, EW, sqrt(VW))
> c1
[1] 2865.435
> c2 <- qnorm(1-0.025, EW, sqrt(VW))
> c2
[1] 3684.565
> W > c2 || W < c1
[1] TRUE
> # ou si l'on considère (W-EW)/sqrt(VW) comme la statistique de test
> c <- qnorm(1-0.025)
> c
[1] 1.959964
> (W-EW)/sqrt(VW) > c
[1] TRUE
```

(g) Calculer la p-valeur sans correction de continuité de Yates.
Rappel : Pour le test bilatéral, la p-valeur est donnée par

$$p = 2min\{P(T \geq t|H_0), P(T \leq t|H_0)\}$$

où T est la statistique de test et t est sa valuer observée avec l'échantillon. Si de plus la fonction de répartition est symétrique par rapport à **zero** sous H0 :

$$p = P(|T| \geq |t||H_0) \tag{0.1}$$

Attention, $W \overset{n\to\infty}{\sim} \mathcal{N}(EW, VW)$, qui est symmetrique mais pas par rapport à **zero**, donc on peut pas utiliser Formule 0.1 avec T donné par W. Donc le calculs

$$p = P(|W| \geq |w||H_0) \tag{0.2}$$

est faux !

Au lieu, on peut considerer en direct $\dfrac{W - EW}{\sqrt{VW}}$ comme le statistique de test, qui suit $\mathcal{N}(0,1)$ asymptotiquement, donc la p-valeur du test est

$$p = P(|\frac{W - EW}{\sqrt{VW}}| \geq |\frac{w - EW}{\sqrt{VW}}||H_0), \tag{0.3}$$

qui est finalement approchée par la normale

$$\begin{aligned} p &\approx P(|Z| \geq \left|\frac{w - EW}{\sqrt{VW}}\right| |Z \sim (0,1)), \\ &\approx 2P(Z \leq -\left|\frac{w - EW}{\sqrt{VW}}\right| |Z \sim (0,1)), \\ &\approx 2(1 - P(Z \leq \left|\frac{w - EW}{\sqrt{VW}}\right| |Z \sim (0,1))). \end{aligned} \tag{0.4}$$

```
> #p-valeur pour le test bitéral
> 2*pnorm(-abs(W-EW)/sqrt(VW))

[1] 1.004501e-06

> # ou
> 2*(1-pnorm(abs(W-EW)/sqrt(VW)))

[1] 1.004501e-06
```

(h) Retrouver tous ces résultats avec R et la fonction wilcox.test.

NB : le test asymptotique effectué dans la fonction wilcox.test utilise la correction de Yates dans le calcul de la p valeur. Quand la taille d'échantillon est assez grand, la correction n'affecte pas trop.

Si avec la correction de continuité de Yates dans Eq (0.4),

$$p = P(|\frac{W - EW}{\sqrt{VW}}| \geq \frac{|w - EW| - 0.5}{\sqrt{VW}}|H_0). \tag{0.5}$$

```
> U = W - n2*(n2+1)/2
> U

[1] 3022

> #Et le test de wilcoxon dans R, applique une correction de continuité de Yates
> 2*(1-pnorm( (abs(W-EW)-0.5)/sqrt(VW)))

[1] 1.016785e-06

> 2*pnorm( (-abs(W-EW)+0.5)/sqrt(VW))
```

```
[1] 1.016785e-06
```

```
> wilcox.test(Y,X,alternative = c("two.sided"),exact=FALSE)

        Wilcoxon rank sum test with continuity correction

data:  Y and X
W = 3022, p-value = 1.017e-06
alternative hypothesis: true location shift is not equal to 0
```

(i) Comparer à un test de Student (test sur les moyennes, modèle gaussien) en supposant les conditions d'applications remplies (indépendance, normalité, homoscédasticité).

Test de Shapiro-Wilk

```
> t.test(X,Y,var.equal = TRUE)

        Two Sample t-test

data:  X and Y
t = -5.4485, df = 128, p-value = 2.508e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.2338976 -0.5764537
sample estimates:
 mean of x  mean of y
0.06374656 0.96892222
```

Exercice 2 **Tests statistiques de Kruskal-Wallis**

Etant donnés les échantillons X et Y de l'exercice 1, nous considérons un autre échantillon d'observations iid

$Z = (7.2, 6.8, 5.6, 5.9, 8.5)$, et on suppose que les 3 échantillons X, Y, Z sont indépendents. On va tester globalement l'homogénéité des trois échantillons comme suit.

(a) Calculer avec R les rangs moyens des échantillons X, Y et Z quand ils sont mélangés, puis calculer la statistique de Kruskal-Wallis.

```
> #On mélange les trois échantillons
> X <- c(4.5, 6.6, 7, 6.7, 3.9, 5.7,5.2)
> Y <- c(4.2, 5.3, 6.4, 5.1, 4.8)
> Z <- c(7.2, 6.8, 5.6, 5.9, 8.5)
> all <- c(X,Y,Z)
> #les rangs, tenant compte de la possibilité des ex-aequos
> rank(all)
 [1]  3 12 15 13  1  9  6  2  7 11  5  4 16 14  8 10 17
> # Calcul des somme de rangs
> n1 <- length(X)
```

```
> n2 <- length(Y)
> n3 <- length(Z)
> n <- n1+n2+n3
> R1 <- sum(rank(all)[1:n1])
> R2 <- sum(rank(all)[(n1+1):(n1+n2)])
> R3 <- sum(rank(all)[(n1+n2+1):(n1+n2+n3)])
> c(R1,R2,R3)
[1] 59 29 65
> #calcul de la statistique de Kruskal Wallis
> KW = 12/(n*(n+1))*( R1^2/n1 + R2^2/n2 +R3^2/n3) - 3*(n+1)
> KW
[1] 5.234734
```

(b) Test exact (alpha=0,05) : comparer la valeur de KW avec une table, voir par exemple si-dessous, et conclure quant le test de Kruskal-Wallis

https://www.dataanalytics.org.uk/critical-values-for-the-kruskal-wallis-test/#grp-6-9

Pour 7, 5 et 5 données, les valeurs critiques sont 5.708 à 5% 7.101 à 2% et 8.108 à 1% ici, on ne rejette donc pas H0 (il faudrait choisir un risque α beaucoup plus grand pour rejetter).

(c) Test asymptotique (alpha=0,05) : calculer la valeur critique avec la loi asymptotique de la statistique de Kruskal-Wallis et conclure le test de Kruskal-Wallis

```
> C <- qchisq(1-0.05,df=2)
> C
[1] 5.991465
> KW > C
[1] FALSE
```

Ne rejette pas H0

(d) Comparer le résultat du test asymptotique avec le test effectué dans R ci-dessous

```
> kruskal.test(list(X,Y,Z))

        Kruskal-Wallis rank sum test

data:  list(X, Y, Z)
Kruskal-Wallis chi-squared = 5.2347, df = 2, p-value = 0.07299
```

La pvaleur de 0.07299 confirme le test calculé à la main, on ne rejette pas H0. Ceci est en fait calculé avec la loi asymptotique du chi-deux à (k-1) degrés de liberté, qu'on peut retrouver dans le calcul de proba suivant :

```
> 1-pchisq(5.2347,df=2)
[1] 0.07299605
```

(e) On considère maintenant les 3 échantillons (indépendants) suivants :

X = (4.5, 6.6, 7, 6.7, 3.9, 5.7, 5.2),

Y =(4.2,5.3,6.4,5.2, 4.8),

Z = (7.7, 7.3, 6.1, 6.4,9.0).

Vérifier que l'on rejette H0 avec la fonction Kruskal-Wallis de R au niveau $\alpha = 0.05$. Ainsi au moins deux des 3 échantillons n'ont pas la même médiane. Que peut-on faire pour préciser cette réponse ? Proposer des idées (sans les effectuer explicitement faute de temps).

```
> X <- c(4.5, 6.6, 7, 6.7, 3.9, 5.7, 5.2)
> Y <- c(4.2, 5.3, 6.4, 5.1, 4.8)
> Z <- c(7.7, 7.3, 6.1, 6.4, 9.0)
> kruskal.test(list(X,Y,Z))

        Kruskal-Wallis rank sum test

data:  list(X, Y, Z)
Kruskal-Wallis chi-squared = 6.3153, df = 2, p-value = 0.04253
```

Pour identifier quelle paire n'a pas de même médiane, on peut faire le test de wilcoxon / mann et whitney sur tous les paire possibles. Et on doit utiliser les approaches (cf. cours) pour contrôler le risque de 1ère espèce.