Licence de Psychologie - TD n° 5

Tests statistiques non paramétriques avec Excel

1 - Test du signe

Ouvrez le fichier W:\Psy3\TD EXCEL\Enfants-PRN.xls.

1.1 -Test unilatéral avec la zone de rejet de H0 "à gauche"

On veut montrer ici que le nombre de différences négatives est significativement grand, ou, de manière symétrique, que le nombre de différences positives est suffisamment faible pour montrer une baisse de l'IDM entre 6 et 24 mois.

On va donc utiliser un test du signe pour comparer les scores des enfants du groupe témoins à 6 mois et à 24 mois. Dans un premier temps, calculons le nombre de différences non nulles et le nombre de différences positives dans la colonne "Différence" (plage D3:D33). On peut pour cela utiliser la fonction NB.SI(). Cette fonction attend deux paramètres : une plage de cellules et une "condition". Elle renvoie en résultat le nombre de cellules de la plage qui satisfont la condition.

Entrez en B35:D39 de la feuille Données PRN-1, le texte et les formules suivantes :

	B	C	D
35	**TEST DU SIGNE**		
36	Nbre de différences non nulles		=NB.SI(D3:D33;"<>0")
37	Nbre de différences positives		=NB.SI(D3:D33;">0")
38	Niveau de significativité		=LOI.BINOMIALE(D37;D36;0,5;VRAI)
39	Valeur Critique à 5%		=CRITERE.LOI.BINOMIALE(D36;0,5;0,05)

Remarquez l'utilisation des fonctions LOI.BINOMIALE et CRITERE.LOI.BINOMIALE.

La fonction LOI.BINOMIALE calcule la fréquence b(n,p,k) (plus simplement b(k)) de la modalité k d'une variable binomiale de paramètres n et p, ou la fréquence cumulée de la valeur k. Ses trois premiers paramètres sont, dans l'ordre, k, n et p. Le quatrième paramètre est une valeur logique.

- Si ce paramètre vaut FAUX, c'est la fréquence b(n,p,k) qui sera calculée.
- Si ce paramètre vaut VRAI, c'est la fréquence cumulée F(k) qui sera calculée.

La fonction CRITERE.LOI.BINOMIALE renvoie la plus petite valeur pour laquelle la distribution binomiale cumulée est supérieure ou égale à une valeur de critère. Autrement dit, elle renvoie la valeur k critique pour laquelle la fréquence cumulée F(k) est supérieure au seuil α choisi, tandis que F(k-1) est inférieure au seuil.

L'ordre des paramètres est n, p, α.

Un petit problème d'intervalles et de piquets :

Ici, la cellule D36 contient la valeur 31, et la cellule D37 contient 12. Ainsi, LOI.BINOMIALE(D37;D36;0,5,VRAI) calcule LOI.BINOMIALE(12;31;0,5;VRAI), c'est-à-dire b(0) + b(1) + ... + b(12) ; c'est le niveau de significativité du résultat (D_+=12) observé.

De même, le résultat fourni par CRITERE.LOI.BINOMIALE(31;0,5,0,05) est la valeur 11, ce qui signifie que :

b(0)+b(1)+...+b(11) > 5%

b(0)+b(1)+...+b(10) < 5%.

La règle de décision précise est donc :

Si D_+ est strictement inférieur à 11, on retient l'hypothèse alternative H_1.

Si D_+ est supérieur ou égal à 11, on retient l'hypothèse nulle H_0.

Vous devriez obtenir comme résultats :

	B	C	D
35	**TEST DU SIGNE**		
36	Nbre de différences non nulles		31
37	Nbre de différences positives		12
38	Niveau de significativité		14,05%
39	Valeur Critique à 5%		11

Conclusion : on n'a pas démontré de différence significative entre l'IDM à 6 mois et l'IDM à 24 mois pour la population d'où a été tiré l'échantillon d'enfants du groupe témoin.

1.2 - Test unilatéral avec la zone de rejet de H0 "à droite"

Il semble plus naturel de raisonner ici sur les différences négatives, et de tenter de montrer que leur nombre est significatif d'une baisse de l'IDM. Il s'agit donc, dans un premier temps, de calculer le nombre de différences non nulles et le nombre de différences négatives dans la colonne "Différence" (plage D3:D33). On utilise comme précédemment la fonction NB.SI().

Entrez en B35:D39 de la feuille Données PRN-2, le texte et les formules suivantes :

	B	C	D
35	**TEST DU SIGNE**		
36	Nbre de différences non nulles		`=NB.SI(D3:D33;"<>0")`
37	Nbre de différences négatives		`=NB.SI(D3:D33;"<0")`
38	Niveau de significativité		`=1- LOI.BINOMIALE(D37-1;D36;0,5;VRAI)`
39	Valeur Critique à 5%		`=D36-CRITERE.LOI.BINOMIALE(D36;0,5;0,05)`

Le problème d'intervalles et de piquets devient plus épineux:

Ici, la cellule D36 contient la valeur 31, et D37-1 vaut 18. Ainsi, LOI.BINOMIALE(D37-1;D36;0,5,VRAI) calcule LOI.BINOMIALE(18;31;0,5;VRAI), c'est-à-dire b(0) + b(1) + ... + b(18).

La formule de la cellule D38 calcule donc b(19)+b(20)+...+b(31) c'est-à-dire le niveau de significativité du résultat (D_-=19) observé.

De même, le résultat fourni par CRITERE.LOI.BINOMIALE(31;0,5,0,05) est la valeur 11, ce qui signifie que :

b(0)+b(1)+...+b(11) > 5%

b(0)+b(1)+...+b(10) < 5%.

En raisonnant sur la queue de la distribution à droite :

b(20) + b(21) .... + b(31) > 5%

b(21) + b(22) + b(31) < 5%

La règle de décision précise est donc :

Si D_- est strictement supérieur à 20, on retient l'hypothèse alternative H_1.

Si D_- est inférieur ou égal à 20, on retient l'hypothèse nulle H_0.

Remarque. La valeur critique peut aussi être obtenue à l'aide de la formule :

```
=CRITERE.LOI.BINOMIALE(D36;0,5;0,95)
```

Vous devriez obtenir comme résultats :

	B	C	D
35	**TEST DU SIGNE**		
36	Nbre de différences non nulles		31
37	Nbre de différences négatives		19
38	Niveau de significativité		14,05%
39	Valeur Critique à 5%		20

Conclusion : on n'a pas démontré de différence significative entre l'IDM à 6 mois et l'IDM à 24 mois pour la population d'où a été tiré l'échantillon d'enfants du groupe témoin.

2 - Test de la médiane

On va utiliser le même fichier W:\Psy3\TD EXCEL\Enfants-PRN.xls et réaliser un test de la médiane pour comparer les IDM à 24 mois des groupes PRN expérimental et PRN témoin.

Si l'on souhaite que les résultats intermédiaires soient calculés par Excel, on pourra utiliser les formules suivantes. Toutes les fonctions utilisées ici sont déjà connues. Remarquer cependant :

- L'utilisation de la fonction MEDIANE avec, comme paramètres, la réunion de deux plages disjointes.

- L'impossibilité d'utiliser une référence de cellule dans le critère de la fonction NB.SI. On est obligé d'indiquer "<=111,5" et non "<=G36".

	F	G	H	I
35	**TEST DE LA MEDIANE pour IDM-24**			
36	Médiane	=MEDIANE(C3:C33;F3:F27)		
37	Observés	PRN Témoin	PRN expérimental	Total
38	<=Médiane	=NB.SI(C3:C33;"<=111,5")	=NB.SI(F3:F27;"<=111,5")	=SOMME(G38:H38)
39	>Médiane	=NB.SI(C3:C33;">111,5")	=NB.SI(F3:F27;">111,5")	=SOMME(G39:H39)
40	Total	=G38+G39	=H38+H39	=SOMME(G40:H40)
41				
42	**Niveau de significativité :**		=TEST.KHIDEUX(G38:H39;L38:M39)	

	K	L	M
37	Théoriques	=G37	=H37
38	=F38	=G$40*$I38/$I$40	=H$40*$I38/$I$40
39	=F39	=G$40*$I39/$I$40	=H$40*$I39/$I$40

Nous devrions ainsi obtenir comme résultats :

TEST DE LA MEDIANE pour IDM-24

Médiane 111.5

Observés	PRN Témoin	PRN expérimental	Total
<=Médiane	19	9	28
>Médiane	12	16	28
Total	31	25	56

Niveau de significativité : 5.988%

Remarque : Le test de la médiane ne met pas en évidence de différence entre les deux groupes. En revanche, un test unilatéral de comparaison de moyennes établit une différence au bénéfice du groupe expérimental. Mais le test de la médiane est moins puissant, et c'est nécessairement un test bilatéral.

3 - Protocoles de rangs

3.1 - La fonction RANG et le calcul des protocoles des rangs

L'utilitaire d'analyse ne comporte pratiquement pas de traitements permettant de faire des tests non paramétriques. Seul l'item Analyse de position permet de déterminer un protocole des rangs, mais sans respecter la convention du rang moyen pour les ex aequo.

La fonction d'Excel permettant de former des protocoles de rangs est la fonction RANG, dont la syntaxe, selon l'aide en ligne, est la suivante :

RANG(nombre;référence;ordre)

nombre est le nombre dont vous voulez connaître le rang.

référence est une matrice, ou une référence à une liste de nombres. Les valeurs non numériques dans référence sont ignorées.

ordre est un numéro qui spécifie comment déterminer le rang de l'argument nombre.

Si l'argument ordre a la valeur 0 (zéro) ou si cet argument est omis, Microsoft Excel calcule le rang d'un nombre comme si la liste définie par l'argument référence était triée par ordre décroissant.

Si la valeur de l'argument ordre est différente de zéro, Microsoft Excel calcule le rang d'un nombre comme si la liste définie par l'argument référence était triée par ordre croissant.

Remarque

La fonction RANG attribue le même rang aux nombres en double. Cependant, la présence de nombres en double affecte le rang des nombres suivants. Par exemple, dans une liste de nombres entiers, si le nombre 10 apparaît deux fois et porte le numéro de rang 5, le nombre 11 se verra attribuer le numéro de rang 7 (aucun nombre n'aura le rang 6).

3.2 - Le calcul du rang moyen

Ouvrez un nouveau classeur Excel et entrez dans les cellules A1:A10 dix scores de sujets, comportant des ex aequo.

Entrez en B1 la formule

```
=RANG(A1;$A$1:$A$10;1)
```

et recopiez jusqu'en B10.

On calcule ainsi le rang des sujets, classés par valeurs croissantes de la variable, en attribuant aux ex aequo le meilleur rang dans leur groupe.

Entrez en C1 la formule

```
=11-RANG(A1;$A$1:$A$10;0)
```

et recopiez jusqu'en C10.

On calcule ainsi le rang des sujets, toujours classés par valeurs croissantes de la variable, mais en attribuant aux ex aequo le moins bon rang dans leur groupe.

Le protocole des rangs, avec la convention du rang moyen pour les ex aequo, peut donc être obtenu de la façon suivante :

Entrez en D1 la formule :

```
=(RANG(A1;$A$1:$A$10;1)+11-RANG(A1;$A$1:$A$10;0))/2
```

et recopiez jusqu'en D10.

Remarque : la constante 11 de cette formule est liée au nombre de données. Elle pourrait être remplacée par l'expression : `NB($A$1:$A$10)+1` ou par une référence à une cellule contenant le nombre d'observations.

4 - Le test de Wilcoxon Mann Whitney

Ouvrez le fichier W:\Psy3\TD EXCEL\Mann-whitney.xls.

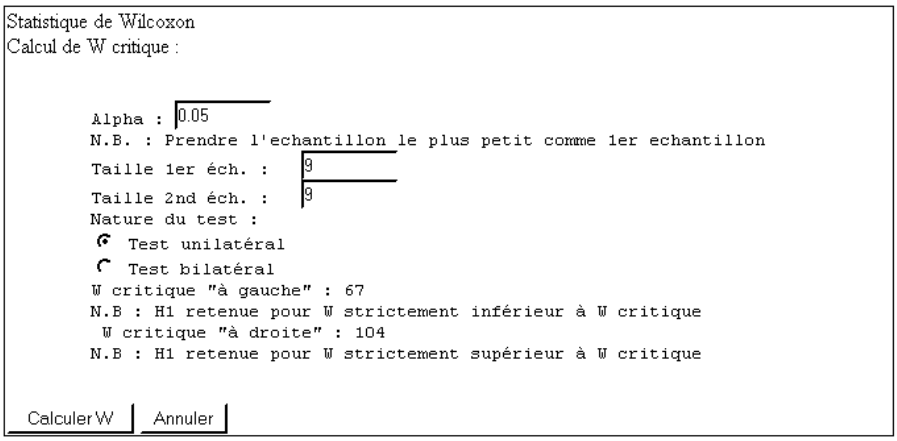
On considère les deux groupes "Maison des parents" et "Famille adoptive".

Calculez en C2:D10 le protocole des rangs pour la réunion des deux groupes, puis en C11 et D11 la somme des rangs dans chacun des deux groupes. Utilisez la table du test de Wilcoxon Mann Whitney pour déterminer si les deux groupes sont homogènes ou non du point de vue de la variable observée.

Les formules utilisées pourraient être :

Rangs Groupe 1	**Rangs Groupe 2**
=(RANG(A2;$A$2:$B$10;1)+19-RANG(A2;$A$2:$B$10;0))/2	=(RANG(B2;$A$2:$B$10;1)+19-RANG(B2;$A$2:$B$10;0))/2
=(RANG(A3;$A$2:$B$10;1)+19-RANG(A3;$A$2:$B$10;0))/2	=(RANG(B3;$A$2:$B$10;1)+19-RANG(B3;$A$2:$B$10;0))/2
=(RANG(A4;$A$2:$B$10;1)+19-RANG(A4;$A$2:$B$10;0))/2	=(RANG(B4;$A$2:$B$10;1)+19-RANG(B4;$A$2:$B$10;0))/2
=(RANG(A5;$A$2:$B$10;1)+19-RANG(A5;$A$2:$B$10;0))/2	=(RANG(B5;$A$2:$B$10;1)+19-RANG(B5;$A$2:$B$10;0))/2
=(RANG(A6;$A$2:$B$10;1)+19-RANG(A6;$A$2:$B$10;0))/2	=(RANG(B6;$A$2:$B$10;1)+19-RANG(B6;$A$2:$B$10;0))/2
=(RANG(A7;$A$2:$B$10;1)+19-RANG(A7;$A$2:$B$10;0))/2	=(RANG(B7;$A$2:$B$10;1)+19-RANG(B7;$A$2:$B$10;0))/2
=(RANG(A8;$A$2:$B$10;1)+19-RANG(A8;$A$2:$B$10;0))/2	=(RANG(B8;$A$2:$B$10;1)+19-RANG(B8;$A$2:$B$10;0))/2
=(RANG(A9;$A$2:$B$10;1)+19-RANG(A9;$A$2:$B$10;0))/2	=(RANG(B9;$A$2:$B$10;1)+19-RANG(B9;$A$2:$B$10;0))/2
=(RANG(A10;$A$2:$B$10;1)+19-RANG(A10;$A$2:$B$10;0))/2	=(RANG(B10;$A$2:$B$10;1)+19-RANG(B10;$A$2:$B$10;0))/2
=SOMME(C2:C10)	=SOMME(D2:D10)
=C11/NB(A2:A10)	=D11/NB(B2:B10)

Pour comparer la somme des rangs obtenue aux valeurs critiques, on peut utiliser les tables fournies en TD ou, si la salle est connectée à l'Internet, utiliser les "tables statistiques en ligne" à l'adresse http://geai.univ-brest.fr/~carpenti/statistiques/table1.php :



On voit que nos résultats conduisent à retenir H0 : on n'a pas mis en évidence de différences entre les deux groupes.

Malgré la faible taille des échantillons, on peut aussi calculer la statistique obtenue en utilisant l'approximation par une loi normale.

Calculez les rangs moyens $\overline{R}_1$ et $\overline{R}_2$ des deux groupes en C12 et D12.

Calculez en G7 et G8 les effectifs N1 et N2 des deux groupes.

Calculez en G9 la valeur du carré de l'erreur type :

$$E^2 = \frac{(n_1 + n_2 + 1)(n_1 + n_2)^2}{12 n_1 n_2}$$

et en G10 la valeur de la statistique de test :

$$Z = \frac{\overline{R}_1 - \overline{R}_2}{E}$$

Indiquez le seuil choisi (5% par exemple) en G11 et la valeur critique obtenue pour un test unilatéral en G12.

On pourra utiliser les formules suivantes (tout en remarquant que la formule donnant l'erreur type peut tout aussi bien être entrée sous la forme =19*18^2/12/9/9):

	F	G
7	Valeur de N1	=NB(A2:A10)
8	Valeur de N2	=NB(B2:B10)
9	E^2	=(G7+G8+1)*(G7+G8)^2/12/G7/G8
10	Z Obs	=(C12-D12)/RACINE(G9)
11	Seuil	0,05
12	Z crit	=LOI.NORMALE.STANDARD.INVERSE(1-G11)

Vous devriez obtenir les résultats suivants :

	Rangs Groupe 1	Rangs Groupe 2
	10	11
	13,5	8,5
	15,5	17
	8,5	18
	1,5	15,5
	4	1,5
	5,5	12
	7	13,5
	3	5,5
Somme Rangs	68,5	102,5
Rang moyen	7,6111	11,3889

Valeur de N1	9
Valeur de N2	9
E^2	6,33
Z Obs	-1,50
Seuil	5,0%
Z crit	1,6449

Exercice : Procéder de même pour effectuer les autres comparaisons de groupes pris deux à deux. La seule comparaison qui nous conduit à accepter l'hypothèse alternative est la troisième : les enfants placés en foyer sont moins souvent absents que les enfants placés en famille adoptive. Utilisez le corrigé (W:\Psy3\TD EXCEL\Mann-whitney-Corrige.xls) pour vérifier vos résultats.

5 - Le test de Wilcoxon

Ouvrez le fichier W:\PSY3\TD EXCEL\Wilcoxon.xls.

Pour utiliser le test de Wilcoxon, il nous faut déterminer le protocole des rangs signés.

Calculer en colonne D la valeur absolue de la différence des scores de l'aîné et du cadet. La fonction "valeur absolue" d'Excel s'appelle ABS.

Calculer en colonne E les rangs de ces différences.

Ces rangs sont reportés en colonne F si la différence est positive et en colonne G si elle est négative.

Ecrire les formules voulue en utilisant la fonction SI.

Par exemple, on pourra obtenir en ligne 2 les formules suivantes :

	D	E	F	G
1	Diff. absolue	Rangs	Rangs +	Rangs -
2	=ABS(B2-C2)	=(RANG(D2;$D$2:$D$21;1)+$J$18+1-RANG(D2;$D$2:$D$21;0))/2	=SI(B2>C2;E2;"")	=SI(B2<C2;E2;"")

Calculez les sommes des rangs des différences positives et des différences négatives en F22 et G22.

Utilisez la table de Wilcoxon pour conclure au seuil de 5%.

Vu la taille de l'échantillon, on peut aussi utiliser l'approximation par une loi normale.

Rappel des formules :

Soit le T maximum des deux sommes de rangs. La statistique de test s'écrit :

$$Z = \frac{T - 0{,}5 - \frac{N(N+1)}{4}}{E} \text{ avec } E^2 = \frac{N(N+1)(2N+1)}{24}$$

Calculez le nombre de différences non nulles en J18.

Calculez le carré de l'erreur type en J19, puis la valeur observée de la statistique de test en J20. Indiquez le seuil choisi et la valeur critique dans les deux cellules suivantes.

On obtiendra par exemple les formules suivantes :

	I	J
18	Diff non nulles	=NB.SI(D2:D21;"<>0")
19	E^2	=J18*(J18+1)*(2*J18+1)/24
20	Z Obs	=(MAX(F22:G22)-0,5-J18*(J18+1)/4)/RACINE(J19)
21	Seuil	0,05
22	Z crit	=LOI.NORMALE.STANDARD.INVERSE(1-J21)

Vous devriez finalement aboutir aux résultats suivants :

E^2	717.5
Z Obs	2.1840
Seuil	0.05
Z crit	1.6449

Exercices. Reprendre le fichier W:\PSY3\TD Excel\Enfants-PRN.xls et traiter les deux comparaisons qui ont été faites à l'aide de tests de rangs.