Licence de Psychologie - TD n° 4

Tests statistiques avec Excel

1 - Test du khi-2

Nous avions vu dans la fiche de TD N°3 comment utiliser Excel pour construire un tableau d'effectifs théoriques et calculer la valeur du χ^2. En complément, nous étudions ici comment faire un test du χ^2 à l'aide d'Excel.

Trois fonctions peuvent être utiles :

```
TEST.KHIDEUX(param1;param2)
LOI.KHIDEUX(param1;param2)
KHIDEUX.INVERSE(param1,param2)
```

1.1 - Mise en oeuvre sur le cas

Ouvrez le classeur W:\PSY3\TD-Excel\Exemple-khi2.xls et consultez l'énoncé donné dans la première feuille.

Affichez la feuille de calcul Test khi-2. Observez le tableau des effectifs théoriques, sous l'hypothèse d'indépendance (plage J6:N8), et rappelez-vous la manière dont il a été calculé. Reportez vous au TD N°3 le cas échéant.

Le niveau de significativité du khi-2 évalué entre les deux tableaux peut alors être obtenu par la formule (que l'on placera par exemple en D11) :

```
=TEST.KHIDEUX(B6:F8;J6:N8)
```

Il est commode d'afficher cette valeur sous forme de pourcentage (ici 8,6%). En TD de statistiques, nous raisonnons plutôt en termes de statistique observée/valeur critique. On peut s'y ramener de la manière suivante :

On choisit un seuil (par exemple 5%), que l'on saisit en G11.

On saisit en D12 la formule :

```
=SI(D11>=G11;"Indépendance";"Dépendance") & " entre les variables"
```

Ainsi, au seuil de 5%, les valeurs observées ne sont pas significatives d'un lien entre les variables. On peut cependant penser que ce résultat décevant est dû au "flou" dans la définition de modalités telles que "absolument opposé" et "opposé" par exemple. Nous allons donc reprendre les mêmes observations, mais en nous limitant maintenant à 3 modalités : désaccord, indifférence, accord.

Complétez le tableau des effectifs observés après regroupement en B19:D21.

Construisez le tableau des effectifs théoriques correspondant en J19:L21.

Calculez ensuite le niveau de significativité du test du khi-2 et la conclusion en D24 et D25.

1.2 - Khi-2 d'ajustement à une loi théorique

Nous avons vu en TD une autre utilisation du test du khi-2 : tester l'ajustement d'une série observée à une loi théorique. Par exemple, on a observé sur un échantillon de 400 sujets la répartition de groupes sanguins suivante :

Groupes sanguins	A	B	AB	O
Observés	192	40	23	145

Cet échantillon peut-il être considéré comme tiré au hasard dans une population où la répartition des groupes sanguins est donnée par :

A	B	AB	O
0,45	0,08	0,03	0,44

Affichez la feuille khi2-ajust.

Calculez en B6:E6 les effectifs théoriques correspondant aux fréquences indiquées en ligne 5.

Calculez le niveau de significativité du test du khi-2 en B9.

1.3 - Méfiance, méfiance...

Une question lancinante dans les manipulations précédentes : Excel calcule-t-il les choses correctement ? Choisit-il le bon nombre de degrés de liberté ? Et l'aide d'Excel est assez muette sur ce dernier point, elle se borne à affirmer (sic) :

> *TEST.KHIDEUX renvoie la valeur de la distribution khi-deux pour la statistique et les degrés de liberté appropriés.*

Affichez la feuille Vérifications. Dans cette feuille, on a fait les calculs complémentaires suivants :

- les contributions au khi-2 en J13:N15 et le khi-2 observés en E16 ;
- le khi-2 critique pour un seuil de 5% et 8 ddl en E15, à l'aide de la formule :

  `=KHIDEUX.INVERSE(0,05;8)`
- le niveau de significativité calculé à partir du khi-2 observé en E17, à l'aide de la formule :

  `=LOI.KHIDEUX(E16;8)`

On peut constater que l'ensemble des résultats est cohérent. On pourrait procéder de même pour l'ajustement à une loi théorique faire la même constatation. Le nombre de degré de liberté utilisé par Excel est donc (l-1)(c-1) si le tableau fourni est un "vrai" rectangle, et (nb observ -1) s'il est uniligne ou unicolonne.

En revanche, l'intervalle de valeurs des paramètres pour lequel une fonction telle que KHIDEUX.INVERSE fournit des résultats corrects est assez limité, comme on pourra le constater sur la feuille Table, dont la structure est analogue, à transposition près, aux tables que nous utilisons en TD. Notez que la version précédente d'Excel affichait une valeur aberrante (50000) et non un message d'erreur lorsque le seuil était trop petit.

2 - Détermination d'un intervalle de confiance

La fonction INTERVALLE.CONFIANCE utilise trois paramètres :

- Le premier paramètre est 1-β, où β est le degré de confiance accordé.
- Le second est l'estimation de l'écart type sur la population.
- Le troisième est la taille de l'échantillon.

Elle renvoie comme résultat la demi-amplitude de l'intervalle de confiance obtenu à l'aide de la loi normale.

Exemple : Nous avons donné en TD de statistiques l'exemple suivant :

Pour un groupe de 500 soldats, le score moyen au test AGCT est de 95 et l'écart type est de 25. Déterminer un intervalle de confiance pour la moyenne, avec le degré de confiance 99%.

On pourra traiter cet exemple en constituant le tableau suivant :

	A	B
1	Moyenne observée	95
2	Ecart type	25
3	Taille échantillon	500
4	Degré de confiance	0,99
5		
6	Demi-amplitude	=INTERVALLE.CONFIANCE(1-B4;B2;B3)
7	Borne inférieure	=B1-B6
8	Borne supérieure	=B1+B6

On retrouve ainsi les résultats obtenus en TD :

	A	B
1	Moyenne observée	95
2	Ecart type	25
3	Taille échantillon	500
4	Degré de confiance	99%
5		
6	Demi-amplitude	2,88
7	Borne inférieure	92,12
8	Borne supérieure	97,88

3 - Comparaisons de moyennes

3.1 - Comparaison de moyennes sur des groupes indépendants ou appareillés : utilisation de la fonction TEST.STUDENT

Ouvrez le classeur Excel W:\PSY3\TD-Excel\Comparaison-Moyennes.XLS.

3.1.1 - Extrait de l'aide d'Excel :

La fonction TEST.STUDENT renvoie la probabilité associée à un test T de Student. Utilisez la fonction TEST.STUDENT pour déterminer dans quelle mesure deux échantillons sont susceptibles de provenir de deux populations sous-jacentes ayant la même moyenne.

Syntaxe

TEST.STUDENT(matrice1; matrice2; uni/bilatéral; type)

matrice1 représente la première série de données.

matrice2 représente la seconde série de données.

uni/bilatéral indique le type de distribution à renvoyer : unilatérale ou bilatérale. Si l'argument uni/bilatéral = 1, la fonction TEST.STUDENT utilise la distribution unilatérale. Si l'argument uni/bilatéral = 2, la fonction TEST.STUDENT utilise la distribution bilatérale.

type représente le type de test t à effectuer.

Si type égale	*Ce test est effectué*
1	*Sur des observations pairées*
2	*Sur deux échantillons de variance égale (homoscédastique)*
3	*Sur deux échantillons de variances différentes (hétéroscédastique)*

3.1.2 - Mise en oeuvre dans le cas "Pédagogie" :

Affichez la feuille de calcul Enoncé-Péda pour vous remettre en mémoire l'énoncé de cet exercice vu en TD de statistiques. Rappelons que cette situation se traite en réalisant un test de comparaison de moyennes sur deux groupes indépendants, à l'aide d'un T de Student.

Affichez la feuille de calcul Pédagogie. Entrez en D19 la formule :

```
=TEST.STUDENT(B5:B14;D5:D14;1;2)
```

Excel nous affiche alors 0,08 ou 8%, c'est-à-dire le niveau de significativité du résultat. Notez qu'il n'existe pas de moyen simple d'obtenir directement la statistique de test, si ce n'est en la recalculant à partir de la valeur précédente. Pour cela, entrez par exemple en cellule D17 la formule :

```
=LOI.STUDENT.INVERSE(D19*2;18)
```

Dans la formule précédente, pourquoi faut-il utiliser D19*2 comme premier paramètre ? Selon les indications fournies par Excel, le premier paramètre de LOI.STUDENT.INVERSE représente la probabilité associée à la loi bilatérale T de Student. Autrement dit, LOI.STUDENT.INVERSE(p, n) renvoie la valeur t telle que : P(X<-t)+P(X>t)=p.

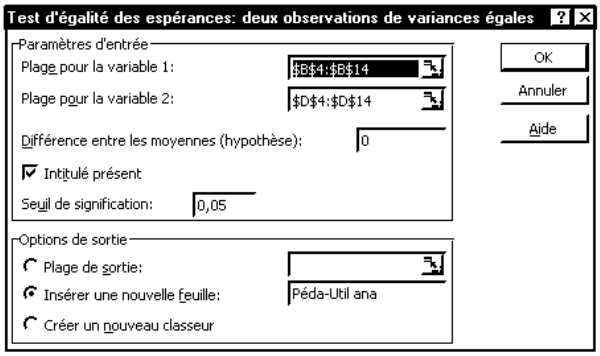
3.1.3 - Mise en oeuvre sur le cas PLPC

Affichez la feuille Enoncé-PLPC. La situation se traite à l'aide d'une comparaison de moyennes sur deux groupes appareillés.

Affichez de même la feuille PL-PC et entrez en C23 une formule analogue à la précédente, mais avec un dernier paramètre égal à 1.

3.2 - Comparaison de moyennes avec l'utilitaire d'analyse

Dans le cas Pédagogie, utilisez l'item Test d'égalité des espérances - Observations de variances égales et complétez la fenêtre de dialogue comme suit :



Les résultats produits peuvent être consultés dans la feuille Péda-Util ana :

Test d'égalité des espérances: deux observations de variances égales		
	Peda1	*Peda2*
Moyenne	3,25	4,25
Variance	2,069444444	2,680555556
Observations	10	10
Variance pondérée	2,375	
Différence hypothétique des moyennes	0	
Degré de liberté	18	
Statistique t	-1,4509525	
P(T<=t) unilatéral	0,081998942	
Valeur critique de t (unilatéral)	1,734063062	
P(T<=t) bilatéral	0,163997884	
Valeur critique de t (bilatéral)	2,100923666	

Dans le cas PL-PC, l'item à utiliser est Test d'égalité des espérances - Observations pairées. La fenêtre de dialogue est analogue à la précédente.

4 - Comparaison de deux proportions

4.1 Cas de deux groupes indépendants

Deux échantillons provenant de deux populations différentes ont passé un test commun.
Dans le premier groupe, d'effectif 150, le taux de succès a atteint 68%.
Dans le deuxième groupe, d'effectif 180, le taux de succès a atteint 55,5%.

Peut-on dire que la seconde population réussit l'épreuve moins facilement que la première ?

Excel ne comporte pas de fonction spécifiquement destinée à traiter ce genre de situation. Notre démarche sera donc très voisine de celle utilisée en calcul manuel.

Ouvrez un nouveau classeur Excel et entrez les données, disposées de la manière suivante :

	A	B	C
1		Groupe 1	Groupe 2
2	Effectif	150	180
3	Taux succès	68%	55.50%

Calculez successivement le taux de succès global en C6, le carré de l'erreur type en C7, la valeur de la statistique Zobs en C8. Le niveau de significativité du résultat peut alors être calculé à l'aide de la fonction LOI.NORMALE.STANDARD() (cf. cellule C9). Rappel des formules :

$$p = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$$

$$Z = \frac{f_1 - f_2}{E} \text{ avec } E^2 = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

On peut aussi raisonner en termes de seuil et de valeur Z critique en utilisant la fonction LOI.NORMALE.STANDARD.INVERSE

	A	B	C
5	**Test de comparaison de proportions (unilatéral)**		
6	Taux de succès global :		`=(B3*B2+C3*C2)/(B2+C2)`
7	Erreur Type au carré :		`=C6*(1-C6)*(1/B2+1/C2)`
8	Erreur Type :		`=RACINE(C7)`
9	Statistique Z observée :		`=(B3-C3)/C8`
10			
11	Niveau de significativité :		`=1-LOI.NORMALE.STANDARD(C9)`
12			
13	Seuil	0.01	
14	Z critique	`=LOI.NORMALE.STANDARD.INVERSE(1-B13)`	

Le résultat du calcul devrait donner :

Test de comparaison de proportions (unilatéral)

Taux de succès global :	61.18%
Erreur Type au carré :	0.00290
Erreur Type :	0.05388
Statistique Z observée :	2.3201
Niveau de significativité :	1.0168%

Seuil	1%
Z critique	2.3263

Remarque : le niveau de significativité peut aussi être obtenu par la formule :

`=1-LOI.NORMALE(B3-C3;0;C8;VRAI)`

Lisez l'article de l'aide d'Excel donnant la signification des paramètres.

Exercice. 1) Faites varier les taux de succès dans les deux groupes. Que devient le résultat du test lorsque les taux varient ?

2) Faites varier les effectifs dans les deux groupes. Avec les taux de succès indiqués, quelles sont les tailles minimales des échantillons permettant d'obtenir un résultat significatif à 5% ?

4.2 Cas de deux groupes appareillés

On veut analyser les résultats aux concours d'entrée de deux écoles d'ingénieurs. Pour cela, on ne considère que les résultats des 300 candidats qui ont présenté ces deux concours à la fois : 60 ont été reçus seulement à l'école A, 44 uniquement à l'école B et 16 aux deux écoles. Peut-on conclure que les deux concours présentent le même niveau de difficulté ?

Ouvrez un nouveau classeur Excel et composez un tableau de contingence rassemblant les données citées dans l'énoncé. Votre tableau pourra avoir l'allure suivante :

	A	B	C	D	E
1			Ecole B		
2			Réussite	Echec	Total
3	Ecole A	Réussite	16	60	
4		Echec	44		
5		Total			300

Calculez ensuite les effectifs qui ne sont pas indiqués par l'énoncé. Par exemple :

Formule : `= C3+D3` en cellule E3

Formule : `= E5-E3` en cellule E4

Formule : `= E4-C4` en cellule D4

Formule : `= C3 + C4` en cellule C5

Formule : `= D3+D4` ou `= E5 - C5` en cellule D4.

Calculez la valeur de la statistique Zobs en C8. Le niveau de significativité du résultat peut alors être calculé à l'aide de la fonction LOI.NORMALE.STANDARD() (cf. cellule C9). Rappel de la formule :

$$Z = \frac{b - c}{\sqrt{b + c}}$$

où b et c désignent les effectifs des deux cases de désaccord.

Vous devriez aboutir aux formules et aux résultats suivants :

	A	B	C
8	Statistique Z observée :		=(C4-D3)/RACINE(C4+D3)
9	Niveau de significativité :		=LOI.NORMALE.STANDARD(C8)

Statistique Z observée :	-1,57
Niveau de significativité :	5,83%

On peut aussi utiliser la statistque du χ^2 de Mac Nemar pour réaliser ce test. Par exemple :

Statistique du khi-2 :		=(C4-D3)^2/(C4+D3)
Niveau de significativité :		=LOI.KHIDEUX(C11;1)

Statistique du khi-2 :	2,46
Niveau de significativité :	11,67%

Remarquez que le niveau de significativité calculé dans le premier cas correspond à un test unilatéral, alors que le second correspond à un test bilatéral (11,67% = 2 * 5.83%).

5 -Exercices de monitorat

Exercice 1

Reprenez le classeur W:\PSY3\TD-EXCEL\MIREAULT.XLS et procédez, à l'aide de l'utilitaire d'analyse, au tri à plat de la variable PVTotal.

Exercice 2

Dans le fichier W:\Psy3\TD-EXCEL\Apprentissage.xls, on donne les scores obtenus par un groupe de 32 sujets à un an d'intervalle.

Calculez l'écart type corrigé des 3 séries de données.

En utilisant un degré de confiance de 95% :

- Déterminez un intervalle de confiance pour la moyenne des scores "avant"
- Déterminez un intervalle de confiance pour la moyenne des scores "après"
- Déterminez un intervalle de confiance pour la différence des scores.

Accepte-t-on l'idée qu'il y a eu un progrès entre les deux époques ?

Exercice 3

2) Ouvrez le classeur W:\Psy3\TD-EXCEL\Enquete-arbres.xls

Une étude par enquête a été réalisée en 1990 sur la représentation de l'arbre d'ornement. Cette étude repose sur l'hypothèse générale d'un lien entre la représentation de l'arbre et l'horizon temporel des sujets.

Quatre terrains d'enquête ont été séléctionnés :

1) Paris XIIè (quartier pauvre en espaces verts)

2) Paris XIVè (quartier riche en espaces verts)

3) Evry (quartier pavillonnaire)

4) Evry (quartier d'immeubles)

1) Dans l'item 2.1 du questionnaire, les sujets interrogés devaient donner leur opinion (accord/désaccord) sur l'affirmation :

L'arbre est le lien entre le passé et l'avenir.

La feuille Lieu habitation donne les tableaux de contingence obtenus en croisant les réponses des sujets d'une part avec leur lieu d'habitation, d'autre part avec leur sexe.

Effectuer des tests du khi-2 pour déterminer au seuil de 5% si :

- Les réponses dépendent du lieu d'habitation.
- Les réponses varient selon le sexe.

Dans les deux cas, on fera apparaître sur la feuille de calcul le niveau de significativité du résultat observé, le seuil choisi et une phrase donnant la conclusion du test.

Alternative : dans le deuxième cas, on n'utilisera pas la fonction TEST.KHI2 ; on constituera le tableau des contributions au khi-2, on calculera le khi-2 observé et le khi-2 critique avant de conclure.

2) La feuille Cyprès donne le tableau de contingence obtenu en croisant les deux variables indiquées pour une photo représentant un cyprès d'Italie.
a) Effectuer un test du khi-2 faisant apparaître sur la feuille de calcul le niveau de significativité du résultat observé, le seuil choisi et une phrase donnant la conclusion du test.
b) On regroupe maintenant les modalités "réponse fausse" et "non réponse". Constituer le tableau des effectifs observés obtenu après regroupement et procéder de nouveau à un test du khi-2.

3) La feuille Noyer donne le tableau de contingence obtenu en croisant les deux variables indiquées pour une photo représentant un noyer commun.
a) Effectuer un test du khi-2 faisant apparaître sur la feuille de calcul le niveau de significativité du résultat observé, le seuil choisi et une phrase donnant la conclusion du test.

b) On regroupe maintenant les modalités "réponse fausse" et "non réponse". Constituer le tableau des effectifs observés obtenu après regroupement. Constituer le tableau des effectifs théoriques et celui des contributions au khi-2. Calculer le khi-2 observé et le khi-2 critique avant de conclure.

4) Dans l'item 2.7 du questionnaire, les sujets interrogés devaient donner leur opinion (accord/désaccord) sur l'affirmation :

Les arbres masquent une partie de la lumière du jour.

On étudie ici si l'opinion du sujet dépend de son âge.
La feuille Item 2-7 donne le tableau de contingence obtenu en croisant les deux variables indiquées.
a) Effectuer un test du khi-2 faisant apparaître sur la feuille de calcul le niveau de significativité du résultat observé, le seuil choisi et une phrase donnant la conclusion du test.
b) On souhaite calculer la fréquence (en pourcentage) de chaque modalité de réponse (accord/désaccord) pour chaque tranche d'âge.
Complétez le tableau de la plage A17:C22 en calculant ces fréquences.

Exercice 4

Reprenez le fichier W:\Psy3\TD-EXCEL\Apprentissage.xls et procédez à une comparaison de moyennes.