EXCEL ET LE TEST DU χ^2

Parmi les nombreuses fonctions statistiques implantées dans EXCEL figure une fonction appelée TEST.KHIDEUX. Cette fonction a-t-elle un rapport avec les tests de Khi deux qui figurent au programme de certaines filières BTSA ? C'est ce que nous allons découvrir dans cet article.
Tout d'abord lisons attentivement les informations fournies dans l'aide en ligne :

> " *Renvoie le test d'indépendance. TEST.KHIDEUX renvoie la valeur de la distribution khi-deux (χ2) pour la statistique et les degrés de liberté appropriés. Utilisez les tests χ2 pour déterminer si les résultats prévus sont vérifiés par une expérimentation.*"

L'aide indique ensuite la syntaxe à utiliser :

> "*TEST.KHIDEUX(plage_réelle;plage_attendue)*
> *plage_réelle représente la plage de données contenant les observations à comparer aux valeurs prévues.*
> *plage_attendue représente la plage de données contenant le rapport du produit des totaux de ligne et de colonne avec le total général.*"

Une remarque donne quelques renseignements supplémentaires :

> "*Le test χ2 calcule d'abord une statistique χ2 puis additionne les différences entre les valeurs réelles et les valeurs prévues. L'équation de cette fonction est TEST.KHIDEUX=p(X>χ2), où :*
>
> $$\chi^2 = \sum_{i=1}^{l}\sum_{j=1}^{c}\frac{\left(A_{ij} - E_{ij}\right)^2}{E_{ij}}$$
>
> *et où :*
> *Aij est la fréquence réelle dans la i-ème ligne et la j-ème colonne.*
> *Eij est la fréquence prévue dans la i-ème ligne et la j-ème colonne.*
> *l est le nombre de lignes.*
> *c est le nombre de colonnes.*
>
> *TEST.KHIDEUX renvoie la probabilité pour une statistique χ2 et des degrés de liberté, où df = (l - 1)(c - 1).*"

Voilà, avec ces explications on est censé être informé et comprendre le fonctionnement de cette fonction KHIDEUX ! ! !
N'oublions pas toutefois le petit exemple qui termine l'aide et qui sert à éclairer ces explications que certains pourraient trouver peu lumineuses :

"*Exemple*

	A	B	C
1	**Réel**		
2		**Hommes**	**Femmes**
3	**D'accord**	58	35
4	**Sans opinion**	11	25
5	**Pas d'accord**	10	23
6			
7	**Prévu**	**Hommes**	**Femmes**
8			
9	**D'accord**	45,35	47,65
10	**Sans opinion**	17,56	18,44
11	**Pas d'accord**	16,09	16,91

La statistique χ2 pour les données ci-dessus est de 16,16957 avec 2 degrés de liberté.
TEST.KHIDEUX(B3:C5,B9:C11) égale 0,000308."

Et bien il va falloir se débrouiller avec ça ! Expliquons donc ce qui précède et tâchons d'être plus clair que l'aide, ce qui a priori ne constitue pas une prouesse pédagogique.

On devine que cette fonction peut être utilisée lorsqu'on veut mettre en place un test d'indépendance.
Dans l'exemple fourni, il semble que l'on souhaite tester l'indépendance des caractères Sexe et Avis dans une population déterminée, à partir d'un échantillon de 162 personnes qui se répartissent, en fonction de leur avis et de leur sexe, comme indiqué dans le tableau intitulé **Réel** (qui constitue, en fait un "*tableau de contingence*").

Le tableau intitulé **Prévu** est le tableau dit des "*effectifs attendus*" ou "*effectifs théoriques*", il est construit sous l'hypothèse

H_0 : "les caractères Sexe et Avis sont stochastiquement indépendants".

L'aide nous rappelle que ce tableau s'obtient, à partir des effectifs marginaux et de l'effectif total du tableau Réel, en effectuant, pour tout i et pour tout j, le calcul : $\frac{n_{i.} * n_{.j}}{n}$ où $n_{i.} = \sum_k n_{ik}$ et $n_{.j} = \sum_k n_{kj}$ sont les effectifs marginaux et n l'effectif total.

En effet, ce tableau donne la répartition des 162 personnes (79 hommes, 83 femmes ; 93 D'accord, 36 Sans opinion et 33 Pas d'accord) lorsqu'on suppose que les caractères Sexe et Avis sont indépendants.
Ceci signifie que pour chaque ligne, c'est-à-dire pour chaque Avis, le pourcentage d'hommes est égal au pourcentage de femmes qui est donc égal au pourcentage de personnes de l'échantillon (sexes confondus) qui ont cet avis.
On doit donc avoir :
$\frac{a}{79} = \frac{d}{83} = \frac{93}{162}$ et la même chose pour les 2 autres lignes. On trouve

	A	B	C	D
1	**Réel**			
2		**Hommes**	**Femmes**	
3	**D'accord**	58	35	93
4	**Sans opinion**	11	25	36
5	**Pas d'accord**	10	23	33
6		79	83	162
7	**Prévu**			
8		**Hommes**	**Femmes**	
9	**D'accord**	a	d	93
10	**Sans opinion**	b	e	36
11	**Pas d'accord**	c	f	33
12		79	83	162
13				

donc : a = $\frac{79\times 93}{162}$, b = $\frac{83\times 93}{162}$ et ainsi de suite.

Avec EXCEL, la construction de ce tableau constitue un très bon exercice d'utilisation des références mixtes. Si, dans la colonne D, on inscrit en D3 la formule =SOMME(B3:C3) que l'on recopie jusqu'en D6 ; si, dans la ligne 6, on inscrit en B6 la formule =SOMME(B3:B5) que l'on recopie en C6 alors, il suffit d'inscrire en B9 la formule =$D3*B$6/$D$6 de la recopier en C9 puis de recopier cette plage jusqu'à la ligne 11 pour obtenir le tableau des effectifs théoriques.

On constate que ce tableau diffère du tableau réel, il s'agit de mesurer "l'écart" entre les deux tableaux pour savoir si cet écart est acceptable (dû aux fluctuations d'échantillonnage) ou bien si cet écart est trop important et nous conduit à refuser l'hypothèse d'indépendance des caractères.
L'instrument de mesure utilisé est la distance du χ^2, elle est obtenue en calculant la somme des carrés des différences des cellules homologues des deux tableaux divisés (ces carrés) par les valeurs des cellules homologues du tableau Prévu. C'est à dire la formule donnée par l'aide en considérant que A_{ij} représente les cellules du tableau Réel et E_{ij} celles du tableau Prévu, ce qui donne ici :

$$\frac{(58-45{,}35)^2}{45{,}35}+\frac{(35-47{,}65)^2}{47{,}65}+\frac{(11-17{,}56)^2}{17{,}56}+\frac{(25-18{,}44)^2}{18{,}44}+\frac{(10-16{,}09)^2}{16{,}09}+\frac{(23-16{,}91)^2}{16{,}91}$$

Si à chaque échantillon de taille 162, on associe le résultat de ce calcul, on obtient une variable aléatoire dont on admet qu'elle suit *approximativement* la loi de χ^2 à (3-1)*(2-1) degrés de liberté (3 et 2 sont les nombres de modalités des 2 caractères étudiés Avis et Sexe).

Effectuons automatiquement le calcul ci-dessus en utilisant la fonction SOMME.XMY2 et les calculs sur les matrices. Une seule formule suffit :
=SOMME.XMY2($B$3:$C$5/RACINE($B$9:$C$11);RACINE($B$9:$C$11)).
Cette formule renvoie la valeur 16,16957507.

Remarques : La fonction SOMME.XMY2, (*somme x moins y au carré*), calcule la somme des carrés des différences entre les éléments de même position de deux matrices (ou plages de cellules) de mêmes dimensions.

Les opérations sur les matrices s'effectuent sur chaque terme des matrices ainsi :
RACINE($B$9:$C$11) renvoie la matrice formée des racines carrées des éléments de la matrice $B$9:$C$11.
$B$3:$C$5/RACINE($B$9:$C$11) divise chaque élément de la plage $B$3:$C$5 par la racine carrée des éléments homologues de la plage $B$9:$C$11.
Si bien que la formule écrite réalise le calcul :

$$\sum\sum\left(\frac{A_{ij}}{\sqrt{E_{ij}}}-\sqrt{E_{ij}}\right)^2$$

, elle fournit donc la valeur du χ^2 calculée sur notre échantillon.

En utilisant la fonction LOI.KHIDEUX(x;degrés_liberté) on obtient la probabilité pour qu'une variable aléatoire suivant la loi de χ^2 à 2 d.d.l. prenne des valeurs supérieures ou égales à la valeur calculée ci-dessus, 16,16957507 :
=LOI.KHIDEUX(SOMME.XMY2($B$3:$C$5/RACINE($B$9:$C$11);RACINE($B$9:$C$11);2))
cette formule renvoie la valeur : 0,000308192.

C'est donc le résultat fourni par la fonction TEST.KHIDEUX !
Ceci signifie que, si les caractères Sexe et Avis sont indépendants, sur 10000 échantillons de 162 personnes il n'y en a que 3 qui donnent une valeur calculée du χ^2 aussi importante.
On peut donc conclure, au vu de cet échantillon, **à une différence de point de vue des hommes et des femmes** au risque de se tromper de 0,03%.

Remarque : *les programmeurs de la fonction TEST.KHIDEUX auraient pu programmer le calcul du tableau "Prévu" et éviter ainsi à l'utilisateur de faire ce calcul. Ceux qui connaissent un peu le VBA pourront reprogrammer la fonction KHIDEUX de telle sorte qu'elle n'ait plus besoin que d'un argument, le tableau Réel.*

En guise d'entraînement, on va corriger un exercice issu d'un sujet de BTSA, il s'agit du sujet Remplacement 1996 France Métropolitaine Options : Productions animales Formation hippique.

Exercice 3

Le tableau ci-dessous donne les résultats concernant la fertilité de trois étalons en insémination artificielle :

Etalons	Nombre de chaleurs fécondées	Nombre de chaleurs Non fécondées
Un atout	20	18
Arthy	20	16
Vainqueur	44	51

Peut-on conclure à une différence de fertilité par chaleur entre ces trois étalons, au seuil de 5%. On utilisera un test du KHI-2.

Eléments de correction en utilisant EXCEL :

Commençons par formaliser le problème. On peut imaginer une population théorique de juments inséminées artificiellement avec la semence de 3 étalons : Un atout, Arthy et Vainqueur. Sur cette population, on étudie les 2 caractères qualitatifs :

- résultat de l'insémination,
- identité du donneur.

Le premier caractère ne possède que 2 modalités : fécondée ou non fécondée (il s'agit de la jument),

Le deuxième caractère possède 3 modalités qui sont les noms des 3 étalons.
L'étude a pour but de conclure à la non indépendance ou à l'indépendance des deux caractères au vu des résultats constatés sur un échantillon aléatoire simple de 169 juments et fournis dans le tableau de contingence ci-dessus.

Pour cela on met en place un test de Khideux :

* Hypothèses :
  H_o : les 2 caractères sont indépendants
  H_1 : les 2 caractères ne sont pas indépendants.

* risque de première espèce : $\alpha = 0,05$.

* Variable aléatoire de décision : X^2 qui associe à chaque échantillon de taille 169 la distance entre le tableau de contingence et le tableau théorique (voir plus haut).
Sous l'hypothèse nulle, cette variable aléatoire suit approximativement la loi de χ^2 à (3-1)(2-1), c'est-à-dire 2, degrés de liberté.

* Règle de décision : Si la valeur du χ^2 calculée à partir de l'échantillon est supérieure ou égale à la valeur lue sur une table des lois de χ^2 pour $\alpha = 0,05$ et 2 d.d.l., alors on rejette l'hypothèse nulle et l'on conclut à la non indépendance des caractères (*ce qui signifie qu'il y a une différence de fertilité par chaleur entre ces trois étalons*). Si la valeur est inférieure il n'y a pas lieu de rejeter l'indépendance des 2 caractères (*ce qui signifie que l'on n'a pas pu mettre en évidence une différence de fertilité entre les trois étalons*).

Nous allons faire les calculs avec EXCEL. On calcule, grâce à la fonction TEST.KHIDEUX, la probabilité pour que, sous l'hypothèse d'indépendance, un échantillon de taille 169 donne un écart entre le tableau constaté et le tableau théorique supérieur à celui que nous obtenons avec notre échantillon.
Si cette probabilité est inférieure à 0,05 nous rejetons l'hypothèse d'indépendance sinon il n'y a pas lieu de la rejeter.
Voici quelle peut être la feuille de calcul :

	A	B	C	D
1	Etalons	Nombre de chaleurs fécondées	Nombre de chaleurs non fécondées	
2	Un atout	20	18	① 38
3	Arthy	20	16	36
4	Vainqueur	44	51	95
5		② 84	85	169
6				
7	Un atout	③ 18,887574	19,112426	
8	Arthy	17,8934911	18,1065089	
9	Vainqueur	47,2189349	47,7810651	
10				
11				
12		TEST.KHIDEUX : ④ 0,5887177		
13				

Formules : D2 : ① =SOMME(B2 :C2)

Recopiée vers le bas jusqu'en D5

B5 : ② =SOMME(B2 :B4)
Recopiée vers la droite jusqu'en C5

B7 : ③ =$D2*B$5/$D$5
Recopiée à droite jusqu'en C7 puis,
le tout recopié vers le bas jusqu'en ligne 9.

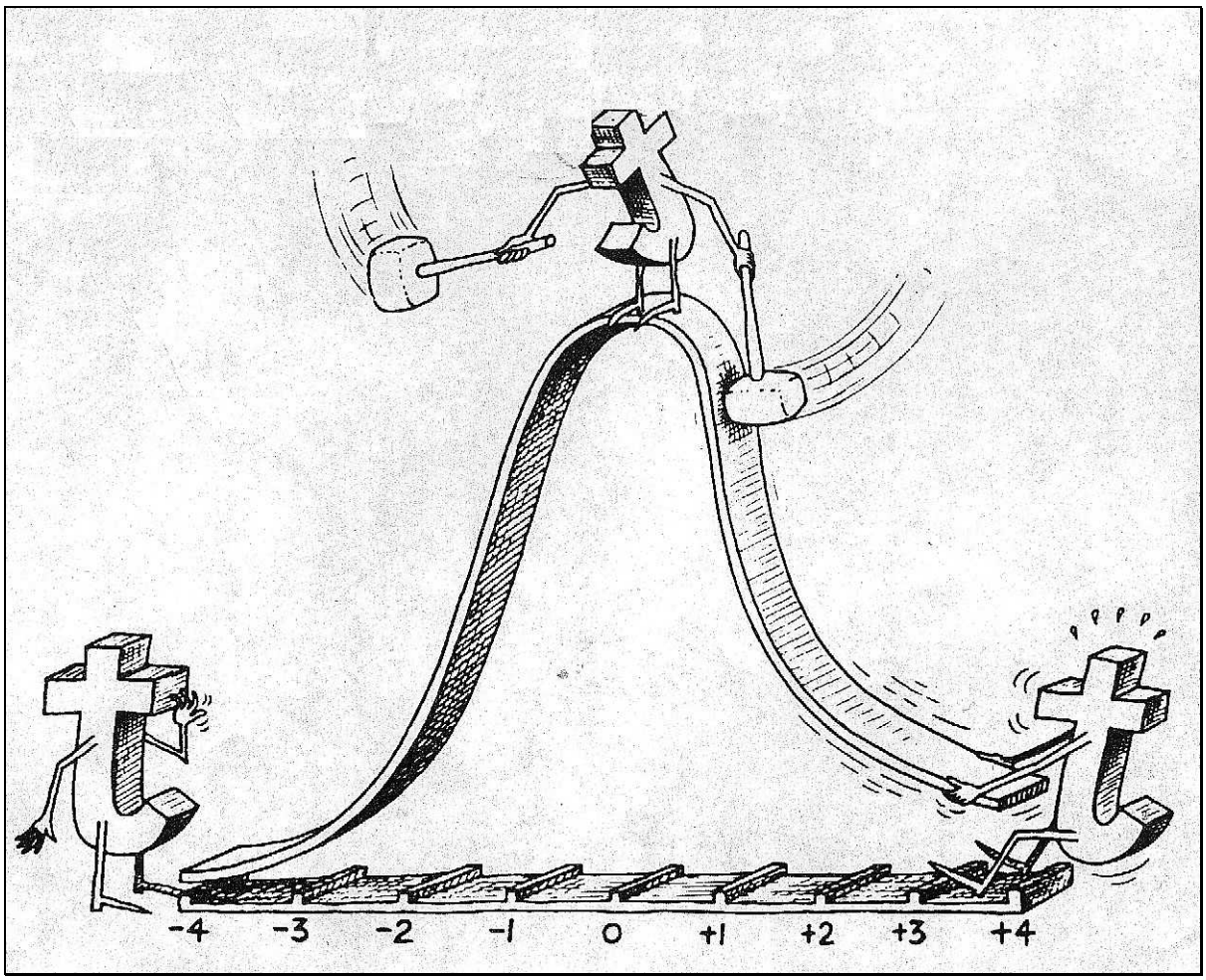
C12 : ④ =TEST.KHIDEUX($B$2:$C$4;$B$7:$C$9)

Notre conclusion : Puisque nous obtenons une probabilité nettement supérieure à 0,05 nous pouvons affirmer que :
au vu de cet échantillon, et au seuil de 5%, il n'y a pas lieu de conclure à une différence de fertilité par chaleur entre ces trois étalons.

Cette même fonction TEST.KHIDEUX peut être utilisée pour mettre en œuvre certains tests d'ajustement. Nous donnerons quelques exemples dans le prochain bulletin.

- = - = - = - = - = - = - = -

La Loi de STUDENT a encore frappé !



Normalisation (*à la soviétique*) d'une loi de Student

(*D'après un dessin illustrant les tables de l'ITCF*)