AUTOUR D'UN TEST DU KHI 2

L'objet de cet article est, au travers d'un exercice somme toute banal, de se poser quelques questions sur nos pratiques et aussi d'apporter quelques réponses.

Voici un texte d'exercice extrait d'un manuel [1] de DEUG, mais qui est tout à fait dans le cadre du programme de certaines filières de BTSA.

Un caractère A est présent ou non pour chaque individu d'une certaine population.
Une hypothèse est émise : la proportion "théorique" des porteurs du caractère A est de 75%.

> *1°) On examine un échantillon de 80 individus. On y trouve 50 porteurs du caractère et 30 non porteurs.*
> ***En utilisant un test du Khi2**, précisez si dans ces conditions, l'hypothèse émise doit être rejetée ou non au risque 5%.*
>
> *2°) Soit un échantillon de 80 individus, on désigne par n le nombre d'individus porteurs du caractère A.*
> *Entre quelles valeurs doit être compris le nombre n pour que l'hypothèse ne soit pas rejetée au risque 5%.*

Une solution rapide et mal rédigée ; mais là n'est pas la question :
Dans la première question, la mise en œuvre du test du Khi2 conduit à décider, au seuil de 5%, le rejet de l'hypothèse (les effectifs théoriques sont respectivement 60 et 20, le Khi2 calculé est de 6,67 et la valeur lue dans la table pour le risque donné est de 3,84).
Dans la deuxième question, les effectifs théoriques sont respectivement n et 80-n, le Khi2 calculé est égal à $\frac{(n-60)^2}{15}$. Il s'agit donc de résoudre l'inéquation $\frac{(n-60)^2}{15} < 3{,}84$ et on trouve que n doit être compris entre 53 inclus et 67 inclus.

Pourquoi ne pas ajouter la question suivante :

> 3°) Compte tenu des moyens de calculs actuels, pouvait-on envisager une autre méthode pour répondre à la question 2 ?

La population est supposée infinie (ou l'échantillon prélevé est supposé simple et aléatoire).Soit X la variable aléatoire qui à chaque échantillon de taille 80 associe le nombre de porteurs du caractère A. Sous l'hypothèse que la proportion de porteurs du caractère A dans la population est de 0,75, a loi de probabilité de X est la loi binomiale de paramètres n = 80 et p = 0,75.
Cette distribution ne pose aucun problème de calcul avec un tableur ou avec une calculatrice un tant soit peu récente.
En répartissant le risque "de façon symétrique en probabilité", le problème consiste donc chercher le plus grand entier a et le plus petit entier b tels que : $P(X \le a) \le 0{,}025$ et $P(X \ge b) \le 0{,}025$.

Ces deux résolutions ne posent aucun problème puisque l'on dispose de l'ensemble des valeurs possibles des probabilités.
On trouve alors a = 51 et b = 68.

Deux questions se posent alors :

Pourquoi ne trouve -t-on pas le même résultat ?

Pouvait-on résoudre la première question avec cette méthode ?

La réponse à la deuxième question, pour un étudiant, est non parce que le texte disait "**En utilisant un test du Khi2**". Par contre, si le texte n'avait pas précisé le type de test à utiliser, alors on aurait très bien pu utiliser cette méthode.
Pour la première question, la réponse est plus facile d'un point de vue théorique, mais plus difficile d'un point de vue pédagogique.
En effet notons D^2 la variable aléatoire de décision utilisée lors de la résolution de la première question, cette variable s'écrit :
$D^2 = \frac{\left(n_1 - 80 \times 0,75\right)^2}{80 \times 0,75} + \frac{\left(n_2 - 80 \times 0,25\right)^2}{80 \times 0,25}$ où n_1 et n_2 désignent respectivement le nombre de porteurs du caractère A et le nombre de non porteurs du caractère.
La notation D^2 est donnée afin d'insister sur l'idée que ce que l'on calcule, c'est un "écart" entre les valeurs d'un tableau et les valeurs correspondantes du tableau que l'on devrait obtenir si l'hypothèse est vraie ; en fait une somme de carrés d'écarts. La variable D^2 est aussi notée K (voir bulletin du GRES N° 6 page 40).
Du fait que n_1 et n_2 sont des entiers positifs de somme égale à 80, la variable D^2 est une variable discrète, (si vous faites les calculs, D^2 ne prend que 61 valeurs distinctes) elle n'est donc sûrement pas distribuée selon une loi du χ^2 à 1 degré de liberté dans le cas de cet exercice. Cette affirmation reste vraie dans le cas où l'on aurait k modalités au lieu de 2.
En résolvant la première et la deuxième question, on a donc utilisé une approximation de la loi de D^2 et ce, sans aucune connaissance (et donc aucune maîtrise) sur cette approximation. D'ailleurs, les manuels eux-mêmes restent très évasifs sur cette question.
Lors de la résolution de la deuxième question, on a montré que $D^2 = \frac{\left(n - 60\right)^2}{15}$, on peut remarquer que $\frac{\left(n - 60\right)^2}{15} = \left(\frac{n - 60}{\sqrt{15}}\right)^2$.
La variable X définie précédemment a pour loi la loi binomiale de paramètres n = 80 et p = 0,75 et donc a pour moyenne np = 60 et pour écart-type $\sigma = \sqrt{15}$. D^2 est donc le carré d'une variable aléatoire de loi normale centrée réduite, là encore il s'agit d'une approximation. Au passage, en posant $U = \frac{\left(n - 60\right)}{\sqrt{15}}$, on voit que l'on aurait très bien pu utiliser la loi normale (centrée réduite) pour faire ce test.

On peut remarquer aussi que si le nombre de porteurs du caractère A est connu, alors le nombre de non porteurs est lui aussi connu. Par suite le test pourrait très bien porter seulement sur le nombre de porteurs du caractère A.
Dans ce cas, la loi binomiale est parfaitement adaptée et de plus cette loi est connue des étudiants (enfin, mieux connue que les lois du Khi 2!).

[1] Mathématiques pour les sciences de la vie P. Trousset et J.F. Morin Mac Graw Hill 1991 (p. 290-291).
[2] Bulletin du GRES n°3 : Quand deux tests se rejoignent (p. 17-19).