Université Mohamed Khider, Biskra 2019/2020
Faculté des Sc. Exactes et Sc. de la Nature et la Vie
Master 1: **M1.4 (Modèle linéaire)**

Série N°1 : test d'indépendance du Khi-deux

Exercice 1 *On interroge* 1587 *étudiants de M2 sur la catégorie socio-professionnelle de leurs parents. Les étudiants suivent différents cursus: écoles d'ingénieurs, écoles de commerce, universités scientifiques, médecine. Les résultats sont les suivants :*

	Ouvriers	*Employés*	*Cadres*	*Professions libérales*
Écoles d'ingénieurs	*50*	*280*	*120*	*20*
Écoles de commerce	*8*	*29*	*210*	*350*
Universités Scientifiques	*150*	*230*	*100*	*40*

On veut étudier l'influence du milieu socio-professionnel des parents sur le type d'étude des enfants.

1-Quelles sont les variables étudiées ? Quelle est leur nature ?

2-On effectue un test d'indépendance du Khi-deux, ou Chi-deux ou noté χ^2*, entre les deux variables*

(a) Préciser les hypothèses nulle et alternative du test.

b) Donner le tableau des fréquences théoriques.

(c) Donner les conditions d'application du test. Sont-elles vérifiées?

(d) Donner la statistique du test du Khi-deux et sa loi sous l'hypothèse nulle.

(e) Vérifier que la valeur observée de la statistique Khi-deux vaut 845.5*.*

(f) Enoncer la règle de décision du test.

(g) La p-valeur associée au test donnée par le logiciel R est p-value $< 2.2 \times 10^{-16}$*. Que pouvez-vous conclure au risque* 5%*?*

Solution.

1. Tout d'abord, il s'agit d'une **table de contingence.** Les deux variables étudiées sont

$$V_1 := \text{(Ouvriers, Employés, Cadres,Professions libérales)}$$

et

$$V_2 := \big(\text{Écoles d'ingénieurs, Écoles de commerce, Universités Scientifiques, Médecine}\big)$$

Leurs nature est **catégorielle** (ou qualitative).

2. a) L'hypothèse nulle et l'hypothèse alternative du test sont respectivement:

H_0 : le milieu socio-professionnel des parents n'influe pas sur le type d'étude des enfants

et

H_1 : le milieu socio-professionnel des parents influe sur le type d'étude des enfants.

b) Le tableau des données est

$$N^* = \begin{pmatrix} 50 & 280 & 120 & 20 \\ 8 & 29 & 210 & 350 \\ 150 & 230 & 100 & 40 \end{pmatrix}$$

Nous avons noté, dans le cours de l'AFC, ce tableau de contingence des observations par

$$N^* = \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{33} \end{pmatrix},$$

où x_{ij} est l'effectif de croisement de deux modalités. Par exemple, l'effectif de croisement des deux modalités

$$\text{Écoles d'ingénieurs} \times \text{Ouvriers}=50=x_{11}.$$

De même

$$\text{Universités Scientifiques} \times \text{Cadres}=100=x_{33}.$$

Le tableau des fréquences observées du nuage est

$$N = \begin{pmatrix} x_{11}/n & x_{12}/n & x_{13}/n & x_{14}/n \\ x_{21}/n & x_{22}/n & x_{23}/n & x_{24}/n \\ x_{31}/n & x_{32}/n & x_{33}/n & x_{33}/n \end{pmatrix} := \begin{pmatrix} f_{11} & f_{12} & f_{13} & f_{14} \\ f_{21} & f_{22} & f_{23} & f_{24} \\ f_{31} & f_{32} & f_{33} & f_{33} \end{pmatrix},$$

avec

$$n = 50 + 280 + 120 + 20 + 8 + 29 + 210 + 350 + 150 + 230 + 140 = 1587,$$

est l'effective totale. Donc

$$N = \begin{pmatrix} 50/1587 & 280/1587 & 120/1587 & 20/1587 \\ 8/1587 & 29/1587 & 210/1587 & 350/1587 \\ 150/1587 & 230/1587 & 100/1587 & 40/1587 \end{pmatrix}.$$

Notons que sous l'hypothèse de dépendance H_1, est la probabilité d'avoir deux modalités croisées $i \times j$, est

$$P\left(V_1 = i, V_2 = j\right) = f_{ij}.$$

Tands dit que sous l'hypothèse d'indépendance H_0, est la probabilité d'avoir deux modalités coisées $i \times j$, est

$$P\left(V_1 = i, V_2 = j\right) = f_{i\cdot} \times f_{\cdot j},$$

où

$$f_{i\cdot} := \sum_{j=1}^{4} f_{ij},\ i = 1, 2, 3 \text{ (fréquences marginales des lignes)},$$

et

$$f_{\cdot j} := \sum_{i=1}^{3} f_{ij},\ j = 1, 2, 3, 4 \text{ (fréquences marginales des colonnes)}.$$

On note par $f_{i\cdot} \times f_{\cdot j}$ **les fréquences théoriques du nuage de points** N^*, qui sont définies par le tableau

$$\widetilde{N} =: \begin{pmatrix} f_{1\cdot} \times f_{\cdot 1} & f_{1\cdot} \times f_{\cdot 2} & f_{1\cdot} \times f_{\cdot 3} & f_{1\cdot} \times f_{\cdot 4} \\ f_{2\cdot} \times f_{\cdot 1} & f_{2\cdot} \times f_{\cdot 2} & f_{2\cdot} \times f_{\cdot 3} & f_{2\cdot} \times f_{\cdot 4} \\ f_{3\cdot} \times f_{\cdot 1} & f_{3\cdot} \times f_{\cdot 2} & f_{3\cdot} \times f_{\cdot 3} & f_{3\cdot} \times f_{\cdot 4} \end{pmatrix}.$$

Dans notre exemple, les fréquences marginales des lignes sont

$$\begin{array}{rrl} f_{1\cdot} = & 50/1587 + 280/1587 + 120/1587 + 20/1587 & = 0.296\,16 \\ f_{2\cdot} = & 8/1587 + 29/1587 + 210/1587 + 350/1587 & = 0.376\,18 \\ f_{3\cdot} = & 150/1587 + 230/1587 + 100/1587 + 40/1587 & = 0.327\,66 \end{array}$$

et fréquences marginales des colonnes sont

$$\begin{array}{rrl} f_{\cdot 1} = & 50/1587 + 8/1587 + 150/1587 & = 0.131\,06 \\ f_{\cdot 2} = & 280/1587 + 29/1587 + 230/1587 & = 0.339\,63 \\ f_{\cdot 3} = & 120/1587 + 210/1587 + 100/1587 & = 0.270\,95 \\ f_{\cdot 4} = & 20/1587 + 350/1587 + 40/1587 & = 0.258\,35 \end{array}.$$

Le tableau des fréquences théoriques du nuage est

$$\widetilde{N} = \begin{pmatrix} 0.296\,16 \times 0.131\,06 & 0.296\,16 \times 0.339\,63 & 0.296\,16 \times 0.270\,95 & 0.296\,16 \times 0.258\,35 \\ 0.376\,18 \times 0.131\,06 & 0.376\,18 \times 0.339\,63 & 0.376\,18 \times 0.270\,95 & 0.376\,18 \times 0.258\,35 \\ 0.327\,66 \times 0.131\,06 & 0.327\,66 \times 0.339\,63 & 0.327\,66 \times 0.270\,95 & 0.327\,66 \times 0.258\,35 \end{pmatrix}.$$

Le calcul donne

$$\widetilde{N} = \begin{pmatrix} 3.881\,5 \times 10^{-2} & 0.100\,58 & 8.024\,5 \times 10^{-2} & 7.651\,3 \times 10^{-2} \\ 4.930\,2 \times 10^{-2} & 0.127\,76 & 0.101\,93 & 9.718\,6 \times 10^{-2} \\ 4.294\,3 \times 10^{-2} & 0.111\,28 & 8.877\,9 \times 10^{-2} & 8.465\,1 \times 10^{-2} \end{pmatrix}.$$

c) La condition d'application du test du Khi-deux est

$$n \times f_{i\cdot} \times f_{\cdot j} \geq 5, \text{ pour tout } i \text{ et } j.$$

Pour vérifier cette condition on doit multiplier $\widetilde{N}$ par $n = 1587$. On a

$$1587\widetilde{N} = 1587 \begin{pmatrix} 3.881\,5 \times 10^{-2} & 0.100\,58 & 8.024\,5 \times 10^{-2} & 7.651\,3 \times 10^{-2} \\ 4.930\,2 \times 10^{-2} & 0.127\,76 & 0.101\,93 & 9.718\,6 \times 10^{-2} \\ 4.294\,3 \times 10^{-2} & 0.111\,28 & 8.877\,9 \times 10^{-2} & 8.465\,1 \times 10^{-2} \end{pmatrix}.$$

Le calcul donne

$$1587\widetilde{N} = \begin{pmatrix} 61.599 & 159.62 & 127.35 & 121.43 \\ 78.242 & 202.76 & 161.76 & 154.23 \\ 68.151 & 176.6 & 140.89 & 134.34 \end{pmatrix}.$$

Il est clair que tout les élements de cette matrice sont supérieurs à 5, donc on peut utiliser le test de Khi-deux.

d) Quand les fréquences $\mathbf{f}_{ij}$ sont alétaoire, la statistique du test Khi-deux est définie par

$$\chi^2 := \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(n\mathbf{f}_{ij} - n\mathbf{f}_{i\cdot}\mathbf{f}_{\cdot j})^2}{n\mathbf{f}_{i\cdot}\mathbf{f}_{\cdot j}},$$

où p est le nombre de lignes et q est nombre de colonnes de la matrice des données N^*. Dans notre exemple $p = 3$ et $q = 4$. Donc

$$\chi^2 := \sum_{i=1}^{3}\sum_{j=1}^{4} \frac{(n\mathbf{f}_{ij} - n\mathbf{f}_{i.}\mathbf{f}_{.j})^2}{n\mathbf{f}_{i.}\mathbf{f}_{.j}}.$$

La loi asymptotique de χ^2, sous l'hypothèse nulle H_0, est la loi de Khi-deux à r degrés de liberté χ^2_r, où

$$r := (p-1)(q-1) = (3-1)(4-1) = 6.$$

En d'autres termes, on peut identifier la loi de χ^2 à celle de χ^2_6.

e) Une fois on observe les fréquences $\mathbf{f}_{ij} = f_{ij}$, alors la valeur observée de la statistique du Khi-deux est

$$\chi^2_{obs} := \sum_{i=1}^{3}\sum_{j=1}^{4} \frac{(nf_{ij} - nf_{i.}f_{.j})^2}{nf_{i.}f_{.j}}.$$

Pour calculer la valeur de χ^2_{obs} on peut utiliser un calculateur scientifique, à savoir les logiciels Matlab ou R. Pour le R, voici les syntaxes à utiliser:

```
Nstart<-matrix(c(50,280,120,20,8,29,210,350,150,230,100,40),ncol=4,byrow=TRUE)
test=chisq.test(Nstart)
```

Résultats du calcul:

X-squared = 845.49, df = 6, p-value < 2.2e-16

Explication:

$$X - squared = \chi^2_{obs} = 845.5,\ df = r = 6,\ p - value < 2.2 \times 10^{-16},$$

où

$$p - value = P\left(\chi^2_6 > \chi^2_{obs}\right) = P\left(\chi^2_6 > 845.5\right) < 2.2 \times 10^{-16}.$$

(f) A un seuil de signification $\alpha \in]0,1[$, la région critique du test est

$$W = \left\{\mathbf{f}_{ij} \in]0,1[: \chi^2_6 > c_\alpha\right\},$$

où c_α est telle que $P\left(\chi^2_6 > c_\alpha\right) = \alpha$. La règle de décision du test est

$$\delta = \begin{cases} 1 & \text{si } \chi^2_6 > c_\alpha \\ 0 & \text{si } \chi^2_6 \leq c_\alpha \end{cases}$$

g) Nous avons ici $\alpha = 5\times10^{-2}$. Comme $P\left(\chi^2_6 > 845.5\right) < 2.2\times10^{-16}$ alors $P\left(\chi^2_6 > 845.5\right) < \alpha$. Comme $P\left(\chi^2_6 > c_\alpha\right) = \alpha$, alors

$$P\left(\chi^2_6 > 845.5\right) < P\left(\chi^2_6 > c_\alpha\right),$$

ce qui implique que

$$P\left(\chi^2_6 \leq 845.5\right) > P\left(\chi^2_6 \leq c_\alpha\right).$$

Le fonction de probabilité est une fonction croissante, alors $\chi^2_{obs} > c_\alpha$ donc $\delta = 1$, c'est à dire on rejette H_0 est on accepte H_1. En conclusion:

Le milieu socio-professionnel des parents influe sur le type d'étude des enfants.