Tableaux de contingence, distributions marginales, distributions conditionnelles, dépendance entre deux variables

BUT-Tech. de Co.
Semestre : 2
A.U. : 2021-2022
Prof. H. El-Otmany

Exercice n°1 On interroge 60 étudiants de BUT-Techniques de Commercialisation de cinq groupes G_i, $1 \leqslant i \leqslant 5$ sur leur niveau d'absentéisme aux cours pendant le semestre 1. Le tableau reporte les résultats où X est la variable correspondante aux groupes et Y est la variable correspondante au niveau d'absentéisme.

Y \ X	G_1	G_2	G_3	G_4	G_5
Rare	7	2	3	6	2
Moyen	3	4	5	1	5
Fréquent	3	5	7	2	5

1. Établir le tableau des distributions en effectifs et en fréquences.

— Le tableau des distributions en effectifs :

Y \ X	G_1	G_2	G_3	G_4	G_5	TOTAL
Rare	7	2	3	6	2	20
Moyen	3	4	5	1	5	18
Fréquent	3	5	7	2	5	12
TOTAL	13	11	15	9	12	60

Comment lire les données du tableau : $n_{11} = 7$ signifie que 7 étudiants qui appartiennent au groupe G_1 et s'absentent rarement.

— Le tableau des distributions en fréquence :

Y \ X	G_1	G_2	G_3	G_4	G_5	TOTAL
Rare	$\frac{7}{60}$	$\frac{2}{60}$	$\frac{3}{60}$	$\frac{6}{60}$	$\frac{2}{60}$	$\frac{20}{60}$
Moyen	$\frac{3}{60}$	$\frac{4}{60}$	$\frac{5}{60}$	$\frac{1}{60}$	$\frac{5}{60}$	$\frac{18}{60}$
Fréquent	$\frac{3}{60}$	$\frac{5}{60}$	$\frac{7}{60}$	$\frac{2}{60}$	$\frac{5}{60}$	$\frac{12}{60}$
TOTAL	$\frac{13}{60}$	$\frac{11}{60}$	$\frac{15}{60}$	$\frac{9}{60}$	$\frac{12}{60}$	$\frac{60}{60} = 1$

Pour calculer la fréquence pour $n = 60$ étudiants, on utilise les formules suivantes

$$f_{ij} = \frac{n_{ij}}{n}$$

où n_{ij} désigne l'individu ayant la modalité i (c'est-à-dire "rare, moyen, fréquent") et la modalité j (c'est-à-dire "G_1, G_2, G_3, G_4, G_5"). On a par exemple

$$f_{11} = \frac{n11}{n} = \frac{7}{60}; \quad f_{12} = \frac{n12}{n} = \frac{2}{60}$$
$$f_{21} = \frac{n21}{n} = \frac{3}{60}; \quad f_{22} = \frac{n22}{n} = \frac{4}{60}$$
$$f_{31} = \frac{n31}{n} = \frac{3}{60}; \quad f_{35} = \frac{n35}{n} = \frac{5}{60}$$

Attention : dans ce cas, la valeur de fréquence en % n'est pas demandée, donc vous n'êtes pas obligés de faire les calculs.

2. Donner la distribution marginale (en effectif et en fréquence) de Xet celle de Y.
   — La distribution marginale en effectif de X :

X	G_1	G_2	G_3	G_4	G_5	TOTAL
Effectif marginale	13	11	15	9	12	60

   et celle en fréquence :

X	G_1	G_2	G_3	G_4	G_5	Total
Fréquence marginale	$\frac{13}{60}$	$\frac{11}{60}$	$\frac{15}{60}$	$\frac{9}{60}$	$\frac{12}{60}$	1

   — La distribution marginale en effectif de Y :

Y	Rare	Moyen	Fréquent	TOTAL
Effectif marginale	20	18	12	60

   et celle en fréquence :

Y	Rare	Moyen	Fréquent	TOTAL
Fréquence marginale	$\frac{20}{60}$	$\frac{18}{60}$	$\frac{12}{60}$	1

3. Calculer la moyenne marginale et la variance marginale de X (resp. de Y).Comme les variables X et Y sont des variables qualitatives, on va coder les variables pour les rendre quantitatives. Cependant, cette moyenne ne donne pas une bonne information sur la base de donnée. Pour remédier à ce problème,il faut représenter les données en utilisant l'analyse par composante principale afin de créer des clusters.
   Attention : chaque codage donne une valeur de moyenne et de variance. Ici la moyenne et la variance ne sont pas unique.

   — En utilisant le codage suivant ($G_1 \longrightarrow 1, G_2 \longrightarrow 2, G_3 \longrightarrow 3, G_4 \longrightarrow 4, G_5 \longrightarrow 5$) pour la variable X et sa distribution marginale en effectif, sa moyenne s'écrit ainsi :

$$\overline{X} = \frac{(1\times 13)+(2\times 11)+(3\times 15)+(4\times 9)+(5\times 12)}{60} = \frac{176}{60} = \frac{44}{15}.$$

   Pour calculer la variance marginale de X, on utilise la formule $V(X) = \overline{X^2} - (\overline{X})^2$ avec

$$\overline{X^2} = \frac{(1^2\times 13)+(2^2\times 11)+(3^2\times 15)+(4^2\times 9)+(5^2\times 12)}{60} = \frac{636}{60} = \frac{53}{5}.$$

   Par conséquent, $V(X) = \frac{53}{5} - \left(\frac{44}{15}\right)^2 = \frac{449}{225} \approx 1.995$.
   — De même, avec le codage (Rare $\longrightarrow 0$, Moyen $\longrightarrow 1$, Fréquent $\longrightarrow 2$) pour la variable Y et sa distribution marginale en effectif, sa moyenne s'écrit ainsi :

$$\overline{Y} = \frac{(0\times 20)+(1\times 18)+(2\times 12)}{60} = \frac{42}{60} = \frac{7}{10}.$$

   On calcule maintenant la variance marginale de Y. On a $V(X) = \overline{Y^2} - (\overline{Y})^2$ avec

$$\overline{Y^2} = \frac{(0^2\times 20)+(1^2\times 18)+(2^2\times 12)}{60} = \frac{66}{60} = \frac{11}{10}.$$

   Par conséquent, $V(Y) = \frac{11}{10} - \left(\frac{7}{10}\right)^2 = \frac{61}{100} \approx 0.61$.
4. Donner les distributions conditionnelles de $X_{|Y=\text{Moyen}}$ et de $Y_{|X=G_3}$.

   — La distribution conditionnelle de $X_{|Y=\text{Moyen}}$ (ici, en effectif) est

X	G_1	G_2	G_3	G_4	G_5
Effectif	3	4	5	1	5

— La distribution conditionnelle de $Y_{|X=G_3}$(ici, en effectif) est

Y	Rare	Moyen	Fréquent
Effectif	3	5	7

5. Déduire si la variable X est indépendante de la variable Y. Commenter le résultat.

Y \ X	G_1	G_2	G_3	G_4	G_5	TOTAL
Rare	$\frac{7}{60}$	$\frac{2}{60}$	$\frac{3}{60}$	$\frac{6}{60}$	$\frac{2}{60}$	$\frac{20}{60}$
Moyen	$\frac{3}{60}$	$\frac{4}{60}$	$\frac{5}{60}$	$\frac{1}{60}$	$\frac{5}{60}$	$\frac{18}{60}$
Fréquent	$\frac{3}{60}$	$\frac{5}{60}$	$\frac{7}{60}$	$\frac{2}{60}$	$\frac{5}{60}$	$\frac{12}{60}$
TOTAL	$\frac{13}{60}$	$\frac{11}{60}$	$\frac{15}{60}$	$\frac{9}{60}$	$\frac{12}{60}$	$\frac{60}{60} = 1$

Pour étudier l'indépendance des variables X et Y, on utilise la formule suivante $F_{ij} = f_i \times g_j$ où f_i désigne la fréquence marginale associée à la modalité de Y_i, g_j désigne la fréquence marginale associée à la modalité de X_j et F_{ij} désigne la fréquence de l'individu ayant la modalité X_i et la modalité Y_j . Prenons par exemple $f_i = \frac{18}{60}$ et $g_j = \frac{15}{60}$. On a

$$f_i \times g_j = \frac{18}{60} \times \frac{15}{60} = \frac{3}{40}$$

Or $F_{ij} = \frac{5}{60} \neq \frac{3}{40}$. On en déduit que les variables aléatoires X et Y sont dépendantes.

Exercice n°2 (Travail personnel) On souhaite étudier la répartition des salaires des employés dans une université en fonction de leurs âges. Le tableau reporte les résultats où X désigne l'âge en années et Y désigne le salaire en euros.

Y(€) \ X (années)	[19; 25[	[25; 35[	[35; 45[	[45; 55[	[55; 65[	TOTAL
[1100; 1300[	130	2	3	6	2	20
[1300; 1500[	109	4	5	1	5	18
[1500; 1700[	56	5	7	2	5	12
[1700; 2000[	25	5	7	2	5	12
[2000; 3000[	0	5	7	2	5	12
[3000; 5000[	0	5	7	2	5	12
TOTAL	13	11	15	9	12	50

Reprendre les questions de l'exercice 1. en modifiant la question 4 (les distributions conditionnelles de $X_{|Y=[1700;2000[}$ et de $Y_{|X=35;45[}$).

Exercice n°3 Une promotion de 150 étudiants en BUT-Techniques de Commercialisation est répartie en cinq groupes G_i, $1 \leqslant i \leqslant 5$. Le tableau reporte les résultats où X est la variable correspondant aux groupes et Y est la variable correspondant à la validation de l'année.

Y \ X	G_1	G_2	G_3	G_4	G_5	TOTAL
V	22	20	24	20	25	111
$\bar{V}$	6	17	5	6	5	39
TOTAL	28	37	28	26	30	150

On souhaite savoir si le fait pour un étudiant de valider son année est indépendante du groupe d'appartenance avec un niveau de confiance 98% (ou seuil de risque 2%)..(Indication : utiliser le test d'indépendance χ^2).

On utilise les étapes de mise en œuvre du test χ^2 (voir le cours) :

1. **Détermination des effectifs observés** O **:** voir le tableau ci-dessus.
2. **Détermination des effectifs théoriques** T **:** on reprend le tableau uniquement avec les effectifs marginaux et on le complète sous l'hypothèse d'indépendance avec la relation $T_{ij} = \frac{n_{i.} \times n_{.j}}{n}$.

X / Y	G_1	G_2	G_3	G_4	G_5	TOTAL
V	20.72	27.38	20.72	19.24	22.20	111
$\bar{V}$	7.28	9.62	7.28	6.76	7.8	39
TOTAL	28	37	28	26	30	150

Exemple de calcul :

$$T_{11} = \frac{n_{1.} \times n_{.1}}{n} = \frac{111 \times 28}{150} \approx 20.72$$
$$T_{21} = \frac{39 \times 28}{150} = 7.28$$
$$T_{15} = \frac{111 \times 30}{150} = 22.20$$
$$T_{25} = \frac{39 \times 30}{150} = 7.8$$

D'après le tableau, les conditions d'application du test χ^2 sont vérifiées : tous les effectifs théoriques sont supérieurs à 5 et l'effectif total est supérieur à 50.

3. **Calcul de l'écart :** on a

$$\chi^2_{obs} = \sum_{i,j} \frac{(O_{ij} - T_{ij})^2}{T_{ij}} = \frac{(22-20.72)^2}{20.72} + \frac{(20-27.38)^2}{27.38} + \frac{(24-20.72)^2}{20.72} + \frac{(20-19.24)^2}{19.24} + \frac{(25-22.20)^2}{22.20}$$
$$+ \frac{(6-7.28)^2}{7.28} + \frac{(17-9.62)^2}{9.62} + \frac{(5-7.28)^2}{7.28} + \frac{(6-6.76)^2}{6.76} + \frac{(5-7.8)^2}{7.8} \approx 10.662$$

4. **Détermination de la valeur du** χ^2 **:**
   — Le tableau de contingence est composé de 2 lignes et 5 colonnes, donc $ddl = (2-1)(5-1) = 4$
   — Pour $\alpha = 98\%$ et $ddl = 4$, la valeur du $\chi^2_{1-\alpha}(ddl)$ dans la table est $\chi^2_{1-\alpha}(ddl) = \chi_{0.02}(4) = 11.668$.
5. **Conclusion sur l'hypothèse d'indépendance :** comme $\chi^2_{obs} < \chi^2_{1-\alpha}(ddl)$, alors on accepte l'hypothèse d'indépendance de X et de Y, et on en déduit que l'appartenance au groupe et la validation de l'année sont indépendantes au seuil de 2%.

Remarque : Pour un niveau de confiance $\alpha = 95\%$, on a $\chi^2_{1-\alpha}(ddl) = \chi_{0.05}(4) = 9.488$. Comme $\chi^2_{obs} > \chi^2_{1-\alpha}(ddl)$, alors on rejette l'hypothèse d'indépendance de X et de Y.