CORRIGÉ

Exercice 6 : *Tableau de contingence et croisement de deux variables quantitatives*

1. **Indiquer les marges du tableau de contingence (tableau 1).**

TABLE 1 – POIDS ET QUANTITÉ DE CHOCOLAT CONSOMMÉE POUR 200 ÉTUDIANTS D'AES AU SEIN DE L'UNIVERSITÉ DE BORDEAUX (MESURANT TOUS 1,75M).

		Poids (en kg)				
		[55 ; 65[	[65 ; 75[	[75 ; 95[	[95 ; 115[	Total
Quantité	[0 ; 50[	27	24	6	3	60
de chocolat consommée	[50 ; 150[	30	30	38	2	100
(*en grammes par semaine*)	[150 ; 450[	1	2	14	23	40
	Total	58	56	58	28	**200**

Source : Inspiré par le Département de Sciences Politiques, Université d'Amsterdam.

Le tableau de contingence nous permet d'aller plus loin que l'étude d'une distribution univariée. Ce type de tableau croise l'information afin d'étudier comment se distribuent les effectifs de chaque modalité d'un caractère suivant les modalités de l'autre, on parle alors de statistique bivariée.

Le tableau de contingence que nous avons déjà analysé dans le premier exercice sur le croisement de deux variables qualitatives peut aussi être utilisé pour croiser deux variables quantitatives traduites en classes. Toutefois, le cas général du croisement de deux variables quantitatives vise à croiser des données ponctuelles (autrement dit en associant chaque individu à une valeur précise de X et de Y).

2. **Calculer et interpréter : n_{34}, f_{12}, $n_{.4}$, $n_{2.}$, $n_{..}$.**

— n_{34} : il s'agit de l'effectif conjoint se situant à l'intersection de la ligne 3 et de la colonne 4, c'est-à-dire l'effectif présentant simultanément la modalité 3 de la variable x et 4 de la variable y.
Ainsi : $n_{34} = 23$
23 étudiants d'AES consomment entre 150 et 450g de chocolat par semaine et pèsent entre 95 et 115kg.

— f_{12} : la fréquence conjointe correspond à la proportion que représente l'effectif conjoint n_{ij} parmi l'effectif total $n_{..}$.
Ainsi, la fréquence conjointe $f_{12} = \frac{n_{12}}{n_{..}} = \frac{24}{200} = 0,12$
12% des étudiants d'AES pèsent entre 65 et 75kg et consomment entre 0 et 50g de chocolat par semaine.

— $n_{.4}$: il s'agit de l'effectif marginal en colonne. Autrement dit, la somme de l'ensemble des effectifs présentant la modalité j, à noter qu'ici i varie et devient donc un ".". Ainsi : $\sum_{j=1}^{q} n_{ij}$.
Alors : $n_{.4} = n_{14} + n_{24} + n_{34} = 3 + 2 + 23 = 28$
28 étudiants pèsent entre 95 et 115kg indépendamment de leur consommation hebdomadaire de chocolat.

— $n_{2.}$: il s'agit de l'effectif marginal en ligne, c'est-à-dire la somme de l'ensemble des effectifs présentant la modalité i, à noter qu'ici j varie et devient donc un ".". Nous avons alors : $\sum_{i=1}^{p} n_{ij}$.
Alors : $n_{2.} = n_{21} + n_{22} + n_{23} + n_{24} = 30 + 30 + 38 + 2 = 100$
100 étudiants consomment entre 50 et 150g de chocolat indépendamment de leur poids.

— $n_{..}$: il s'agit de l'effectif total (la population étudiée), autrement dit, la somme de l'ensemble des effectifs conjoints ou la somme des effectifs marginaux en ligne comme en colonne.
On a : $n_{..} = \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij} = \sum_{j=1}^{q} n_{.j} = \sum_{i=1}^{p} n_{i.}$
Ici $n_{..} = 27 + 24 + 6 + 3 + 30 + 30 + 38 + 2 + 1 + 2 + 14 + 23 = 58 + 56 + 58 + 28 = 60 + 100 + 40 = 200$

3. **Donner la signification des profils lignes et des profils colonnes.**
   — On notera la fréquence conditionnelle $f_{i/j}$ ou f_i^j, c'est-à-dire la fréquence de $x = x_i$, sous condition de $y = y_j$. Ainsi, les profils lignes et colonnes indiquent pour une valeur x_i ou y_j l'ensemble des valeurs de x_i ou y_j.
   Ainsi, les profils en ligne et en colonne correspondent à l'ensemble des fréquences conditionnelles, déterminées par le rapport de l'effectif conjoint sur l'effectif marginal : $f_{i/j} = \frac{n_{ij}}{n_{.j}}$ (en colonne) et $f_{j/i} = \frac{n_{ij}}{n_{i.}}$ (en ligne).

4. **Calculer et indiquer la signification de : $f_{i=2/j=3}$, $f_{j=2/i=3}$.**
   — $f_{i=2/j=3} = \frac{n_{23}}{n_{.3}} = \frac{38}{58} = 0,66$
   Parmi les étudiants pesant entre 75 et 95kg, 66% consomment entre 50 et 150g de chocolat par semaine.
   — $f_{j=2/i=3} = \frac{n_{32}}{n_{3.}} = \frac{2}{40} = 0,05$
   Parmi les étudiants qui consomment entre 150 et 450g de chocolat par semaine, 5% pèsent entre 65 et 75kg.

5. **Calculer et interpréter les moyennes conditionnelles de Y.**
   Soit $\bar{x}_j$ la moyenne conditionnelle de x pour $y = y_j$, il y a donc q distributions conditionnelles de x selon y, et inversement, p distributions conditionnelles de y selon x.
   Ainsi : $\bar{x}_j = \frac{1}{n_{.j}} \sum_{i=1}^{p} n_{ij} x_i$ avec $1 \leq j \leq q$
   Et : $\bar{y}_i = \frac{1}{n_{i.}} \sum_{j=1}^{q} n_{ij} y_j$ avec $1 \leq i \leq p$

   Dans notre cas, pour la variable de poids (Y) :
   $\bar{y}_1 = \frac{1}{60}[(27.60) + (24.70) + (6.85) + (3.105)] = \mathbf{68{,}75}$
   $\bar{y}_2 = \frac{1}{100}[(30.60) + (30.70) + (38.85) + (2.105)] = \mathbf{73{,}40}$
   $\bar{y}_3 = \frac{1}{40}[(1.60) + (2.70) + (14.85) + (23.105)] = \mathbf{95{,}13}$

   Les étudiants consommant entre 0 et 50g de chocolat par semaine pèsent en moyenne 68,75kg.
   Les étudiants consommant entre 50 et 150g de chocolat par semaine pèsent en moyenne 73,40kg.
   Les étudiants consommant entre 150 et 450g de chocolat par semaine pèsent en moyenne 95,13kg.

6. **Calculer les variances conditionnelles de y.**
   De la même manière que les moyennes conditionnelles : $V_i(y) = \frac{1}{n_{i.}} \sum_{j=1}^{q} n_{ij}(y_j - \bar{y}_i)^2$

   — $V_1(y) = \frac{1}{60}[27.(60-76,35)^2 + 24.(70-76,35)^2 + 6.(85-76,35)^2 + 3.(105-76,35)^2] = \mathbf{184{,}95}$

   Et : $\sigma_{1y} = \sqrt{184,95} = \mathbf{13{,}60}$
   L'écart moyen entre le poids des étudiants consommant entre 0 et 50g de chocolat par semaine et le poids moyen de ces mêmes étudiants est de 13,60kg.

   — $V_2(y) = \frac{1}{100}[30.(60-76,35)^2 + 30.(70-76,35)^2 + 38.(85-76,35)^2 + 2.(105-76,35)^2] = \mathbf{137{,}14}$

   Et : $\sigma_{2y} = \sqrt{137,14} = \mathbf{11{,}71}$
   L'écart moyen entre le poids des étudiants consommant entre 50 et 150g de chocolat par semaine et le poids moyen de ces mêmes étudiants est de 11,71kg.

   — $V_3(y) = \frac{1}{40}[1.(60-76,35)^2 + 2.(70-76,35)^2 + 14.(85-76,35)^2 + 23.(105-76,35)^2] = \mathbf{506{,}86}$

   Et : $\sigma_{3y} = \sqrt{506,86} = \mathbf{22{,}51}$

L'écart moyen entre le poids des étudiants consommant entre 150 et 450g de chocolat par semaine et le poids moyen de ces mêmes étudiants est de 22,51kg.

Bonus :

Dans quelle catégorie (consommation hebdomadaire de chocolat) le poids des étudiant est-il le plus homogène ?

A l'instar de la question 4 de l'exercice précédent, il faut utiliser les coefficients de variation afin de pouvoir comparer la dispersion des distributions conditionnelles de y.

Ainsi :
$CV_{1y} = \frac{\sigma_{1y}}{\overline{y1}} = \frac{13{,}60}{68{,}75} = 0,20$
$CV_{2y} = \frac{\sigma_{2y}}{\overline{y2}} = \frac{11{,}71}{73{,}40} = 0,16$
$CV_{3y} = \frac{\sigma_{3y}}{\overline{y3}} = \frac{22{,}51}{95{,}13} = 0,24$

Nous avons donc : $CV_{2y} < CV_{1y} < CV_{3y}$

7. **Calculer et interpréter les moyennes marginales.**
La moyenne marginale correspond à la moyenne de la distribution marginale de x ou de y, c'est-à-dire :
$\overline{x} = \frac{1}{n_{..}} \sum_{i=1}^{p} n_{i.} x_i$ avec comme toujours x_i le centre de classe.
Aussi : $\overline{x} = \frac{1}{n_{..}} \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij} x_i$
Dans notre cas :
$\overline{x} = \frac{1}{n_{..}}(n_{1.}x_1 + n_{2.}x_2 + n_{3.}x_3) = \frac{(25.60)+(100.100)+(300.40)}{200} = 117,5$
C'EST EXACTEMENT LA MOYENNE ARITHMÉTIQUE CALCULEÉ DANS L'EXERCICE PRÉCÉDENT.

$\overline{y} = \frac{1}{n_{..}}(n_{.1}y_1 + n_{.2}y_2 + n_{.3}y_3 + n_{.4}y_4) = \frac{(58.60)+(56.70)+(58.85)+(28.105)}{200} = 76,35$
En moyenne les étudiants d'AES interrogés pèsent 76,35kg.

8. **Démontrer que :**

$$\overline{x} = \frac{1}{n_{..}} \sum_{j=1}^{q} n_{.j} \overline{x_j}$$

Nous savons par définition que :

$\overline{x} = \frac{1}{n_{..}} \sum_{i=1}^{i=p} n_{i.} x_i = \frac{1}{n_{..}} \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij} x_i$

D'après la propriété de commutativité de $\sum$:

$\frac{1}{n_{..}} \sum_{j=1}^{q} [\sum_{i=1}^{p} n_{ij} x_i]$

Or :

$\overline{x}_j = \frac{1}{n_{.j}} \sum_{i=1}^{p} n_{ij} x_i$

Autrement dit :

$\sum_{i=1}^{p} n_{ij} x_i = \overline{x}_j n_{.j}$

On a bien :

$\overline{x} = \frac{1}{n_{..}} \sum_{j=1}^{q} n_{.j} \overline{x_j}$

9. **Uniquement en observant le tableau de contingence, quelle pourrait être la relation entre la quantité de chocolat ingérée et le poids des étudiants ?**

   Lorsque l'on met en relation deux variables c'est pour les comparer, par intuition on pense donc qu'elles on un certain rapport entre elles.
   Dans le cas qui nous occupe ici, il semble que plus les étudiants mangent du chocolat et plus ils ont tendance à avoir un poids élevé (*conf.* les moyennes conditionnelles). Toutefois l'inverse semble également vrai, plus les étudiants pèsent lourds et plus ils ont tendance à manger du chocolat...
   Ainsi, il est clair qu'une relation existe, mais il est impossible dans ce cadre de conclure quant à la causalité de X sur Y et inversement. Cette question sera abordée plus précisément dans les exercices suivants (exercices 7, 8 et 10).

`Corrigé [en ligne le : 16/10/2016]` `[006A]`