Chapitre 11

Relation entre deux variables qualitatives

Après avoir vu la puissance des liaisons entre variables quantitatives, il peut paraître étrange de s'intéresser à la relation entre deux variables qualitatives. Pourtant, cet aspect est fondamental, car il existe beaucoup plus de variables qualitatives que de variables quantitatives. Par ailleurs, toute variable quantitative peut être analysée par la liaison faisant lobjet de ce chapitre.

11.1 Corrélation entre deux variables qualitatives

Soient X et Y deux variables qualitatives prenant respectivement k et l modalités, notées x_i et y_j. On dresse un tableau de contingence (Fig. 11.1). Pour mesurer une corrélation entre des variables qualitatives, on calcule le coefficient d^2 :

$$d^2 = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} = n\sum_{i=1}^{k}\sum_{j=1}^{l}\left(\frac{{n_{ij}}^2}{n_{i.}n_{.j}} - 1\right) = n\sum_{i=1}^{k}\sum_{j=1}^{l}\left(\frac{{f_{ij}}^2}{f_{i.}f_{.j}} - 1\right) \tag{11.1}$$

avec n l'effectif total, $n_{i.}$, $n_{.j}$, n_{ij} les effectifs en ligne du caractère X, en colonne du caractère Y et représentant le caractère x_i pour le facteur X et le caractère y_j pour le facteur Y. Plus d^2 est petit, plus la liaison entre les variables X et Y est forte. d^2 mesure l'écart à l'indépendance. L'**indépendance parfaite** implique que la valeur de d^2 soit nulle. Pour mesurer l'écart à l'indépendance, il faut trouver la borne supérieure de d^2. Comme $\frac{n_{ij}}{n_{i.}} \leq 1$ et $\frac{n_{ij}}{n_{.j}} \leq 1$, on obtient facilement que $\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{{n_{ij}}^2}{n_{i.}n_{.j}} \leq \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{n_{ij}}{n_{.j}}$ avec $\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{n_{ij}}{n_{.j}} = l$, donc $d^2 \leq n\,(l-1)$, et de la même façon $d^2 \leq n\,(k-1)$. On en conclut que :

$$d^2 \leq \inf(k-1, l-1) \tag{11.2}$$

Si la borne est atteinte, il existe une relation fonctionnelle entre les variables X et Y. En effet, en supposant, par exemple, $l < k$, on obtient :

$$d^2 = n(l-1) \tag{11.3}$$

si $\frac{n_{ij}}{n_{i.}} = 1$, ou, en d'autres termes, il n'y a aucune case nulle.

	y_1	y_2	...	y_j	...	y_l	**Total**
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1l}	$n_{1.}$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2l}	$n_{2.}$
...	...	...	...	...	...	...	...
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{il}	$n_{i.}$
...	...	...	...	...	...	...	...
x_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{kl}	$n_{k.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.l}$	n

TABLE 11.1 – Tableau de contingence entre deux variables qualitatives

11.1.1 Le coefficient de contingence

$$\kappa = \sqrt{\frac{d^2}{d^2 + n}} \tag{11.4}$$

11.1.2 Le coefficient de Pearson

$$\phi^2 = \frac{d^2}{n} \tag{11.5}$$

11.1.3 Le coefficient de Tschuprov

Alexander Alexandrovich Chuprov (ou Tschuprov) (1874-1926)

$$T = \frac{\phi^2}{\sqrt{(p-1)(q-1)}} \tag{11.6}$$

avec ϕ^2 est le coefficient de Pearson.

11.2 Test d'indépendance de deux variables qualitatives (ou test d'indépendance du χ^2)

L'étude de la distribution d^2 permet de tester l'indépendance, ou plus exactement, de rejeter cette hypothèse. Ainsi, la liaison entre deux variables qualitatives peut être établie.

Dans une population P, chaque individu possède deux caractères qualitatifs X et Y ayant les modalités respectives $x_1, \ldots, x_k$ et $y_1, \ldots, y_l$. Pour tout $i \in \{1, 2, \ldots, k\}$ et $j \in \{1, 2, \ldots, l\}$, on connaît le nombre O_{ij} d'individus présentant les modalités x_i et y_j. On note $n = \sum_{j=1}^{l} \sum_{i=1}^{k} O_{ij}$ l'effectif total de l'échantillon étudié.

L'hypothèse nulle h_0 teste si les deux caractères X et Y sont indépendants. La variable Y est statistiquement indépendante de la variable X si les distributions conditionnelles de Y à X fixées sont identiques, c'est-à-dire si leur fréquences marginales sont identiques. De même, la variable X est statistiquement indépendante de la variable Y si les distributions conditionnelles de X à Y fixé sont identiques

Premièrement, on calcule les effectifs théoriques sous H_0. C_{ij} est l'effectif des individus présentant les modalités x_i et y_j si l'hypothèse H_0 était vérifiée. On note les effectifs marginaux :

$$T_j = \sum_{i=1}^{k} O_{ij} \tag{11.7}$$

et

$$S_i = \sum_{j=1}^{l} O_{ij} \tag{11.8}$$

Sous H_0, les événements x_i et y_j sont indépendants :

$$\Pr(x_i \cap y_j) = \Pr(x_i) \Pr(y_j) \tag{11.9}$$

c'est-à-dire

$$\frac{C_{ij}}{n} = \frac{S_i}{n} \times \frac{T_j}{n} \tag{11.10}$$

On a alors $C_{ij} = \frac{S_i T_j}{n}$ et les calculs se présentent comme dans le cas du test χ^2 d'homogénéité.

Secondement, grâce au théorème, sous H_0, la variable aléatoire Y prenant sur chaque échantillon de taille n la valeur :

$$\chi^2{}_C = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(O_{ij} - C_{ij})^2}{C_{ij}} \tag{11.11}$$

suit une loi du χ^2 à $v = (k-1)(l-1) = kl - 1$ degrés de liberté.

Remarque. Le calcul devient très simple si $k = l = 2$.

En général, on exige que $C_{ij} \geq 5$ pour tout i et pour tout j. Si ce n'est pas le cas, on effectuera des regroupements.

Le risque de première espèce α étant fixé et v étant connu, on lit dans les tables la valeur $\chi^2{}_C$ telle que $\Pr(Y \geq \chi^2{}_C) = \alpha$.

— Si $\chi^2{}_C \geq \chi^2{}_\alpha$, H_0 est rejetée au risque α, c'est-à-dire que d^2 est supérieur à ${d_\alpha}^2$.

— Si $\chi^2{}_C < \chi^2{}_\alpha$, H_0 ne peut pas être rejetée.

Remarque importante. Le test d'indépendance du χ^2, au vu de sa nature, fonctionne également avec les variables quantitatives.

Exemple On dispose de deux traitements A et B contre une maladie M. On souhaite évaluer si la nature du traitement influe sur la guérison des personnes ayant contracté cette maladie.

L'équipe médicale d'un hôpital a étudié les statistiques concernant les 281 personnes affectées par la maladie M qu'elle a accueillies au cours du dernier trimestre.

— Sur les 173 personnes ayant reçu le traitement A, 139 ont été guéries au bout de cinq jours de traitement.

— Sur les 108 personnes ayant reçu le traitement B, 98 ont été guéries au bout de cinq jours de traitement.

La guérison au bout de cinq jours est-elle liée à la nature du traitement reçu ?

On pose deux variables qualitatives : l'état du patient au bout de cinq jours décrit par deux modalités : « guéri » ou « non guéri » et la nature du traitement (A ou B). On obtient un tableau de contingence 11.2.

On pose les deux hypothèses :

1. H_0 : la guérison au bout de cinq jours de la maladie ne dépend pas du traitement suivi. Les deux critères sont indépendantes.
2. H_1 : la guérison au bout de cinq jours de la maladie dépend du traitement suivi. Les deux critères sont dépendantes.

		Variable dépendante				Somme marginale par ligne
		Guéri ($j=1$)		Non guéri ($j=2$)		
		Effectif observé	Effectif théorique	Effectif observé	Effectif théorique	
Variable indépendante	Traitement A $i=1$	139	145,91	34	27,09	173
		$\chi_{11}{}^2 = \frac{(139-145,91)^2}{145,91} = 0,33$		$\chi_{12}{}^2 = \frac{(34-27,09)^2}{27,09} = 1,76$		
	Traitement B $i=2$	98	91,09	10	16,91	108
		$\chi_{21}{}^2 = \frac{(139-91,09)^2}{91,09} = 0,58$		$\chi_{22}{}^2 = \frac{(10-16,91)^2}{16,91} = 2,82$		
Somme marginale par colonne		237		44		281

TABLE 11.2 – Tableau de contingence et valeurs du χ^2

Dans un test d'indépendance, c'est le rejet de l'hypothèse nulle qui permet de mettre en évidence une liaison entre deux variables.

On peut alors construire le tableau de contingence (Tab.11.2), auquel on ajoute le calcul des effectifs théoriques attendus C_{ij} sous H_0 en utilisant les fréquences théorique en se servant des sommes marginales des colonnes :

$$\begin{array}{ll} f_{11} = \frac{n_{.1}}{n} & f_{12} = \frac{n_{.2}}{n} \\ f_{21} = \frac{n_{.1}}{n} & f_{22} = \frac{n_{.2}}{n} \end{array} \tag{11.12}$$

tandis que $C_{ij} = n_{i.} f_{ij}$. On peut alors calculer la valeur du χ^2 par cellule ${\chi_{ij}}^2$, et pour tout le tableau $\chi^2 = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} {\chi_{ij}}^2 = 5,44$.

Il est à noter que l'écart à l'indépendance le plus important ${\chi_{22}}^2$ correspond à la plus forte contribution par rapport à la liaison des deux variables.

Il ne reste qu'à déterminer la p_{value} de $\upsilon = (n_1 - 1)(n_2 - 1)$ degrés de liberté, c'est-à-dire $\upsilon = 1$. Le test est unilatéral. Pour un seuil de confiance α, il existe une table spécifique. Pour $\alpha = 0,05$, ${\chi_C}^2 = 3,841$. Ici, $\chi^2 < {\chi_C}^2$, ce qui signifie que H_0 est rejetée. Le taux de guérison entre les deux traitements est différent. Les deux variables sont liées. En cas d'indépendance, il y aurait eu moins de 2,5 % de chances d'obtenir de telles différences entre les deux traitements. On aurait eu $0,01 < p_{value} < 0,025$.

Bibliographie

[Wonnacott et Wonnacott, 1995] WONNACOTT, T. H. et WONNACOTT, R. J. (1995). Statistique. Économie – Gestion – Sciences – Médecine. Économica, Paris.