Statistique descriptive bidimensionnelle

Résumé

Liaisons entre variables quantitatives (corrélation et nuages de points), qualitatives (contingence, mosaïque) et de types différents (rapport de corrélation). Introduction au cas multidimensionnel.

Retour au plan.

1 Introduction

Dans cette section, on s'intéresse à l'étude simultanée de deux variables X et Y, étudiées sur le même échantillon, toujours noté Ω. L'objectif essentiel des méthodes présentées est de mettre en évidence une éventuelle variation simultanée des deux variables, que nous appellerons alors *liaison*. Dans certains cas, cette liaison peut être considérée *a priori* comme *causale*, une variable X expliquant l'autre Y ; dans d'autres, ce n'est pas le cas, et les deux variables jouent des rôles symétriques. Dans la pratique, il conviendra de bien différencier les deux situations et une liaison n'entraîne pas nécessairement une causalité. Sont ainsi introduites les notions de covariance, coefficient de corrélation linéaire, régression linéaire, rapport de corrélation, indice de concentration, khi-deux et autres indicateurs qui lui sont liés. De même, nous présentons les graphiques illustrant les liaisons entre variables : nuage de points (*scatter-plot*), diagrammes-boîtes parallèles, diagramme de profils, tableau de nuages (*scatter-plot matrix*).

2 Deux variables quantitatives

2.1 Nuage de points

Il s'agit d'un graphique très commode pour représenter les observations simultanées de deux variables quantitatives. Il consiste à considérer deux axes perpendiculaires, l'axe horizontal représentant la variable X et l'axe vertical la variable Y, puis à représenter chaque individu observé par les coordonnées des valeurs observées. L'ensemble de ces points donne en général une idée assez bonne de la variation conjointe des deux variables et est appelé *nuage*. On notera qu'on rencontre parfois la terminologie de *diagramme de dispersion*, traduction plus fidèle de l'anglais *scatter-plot*.

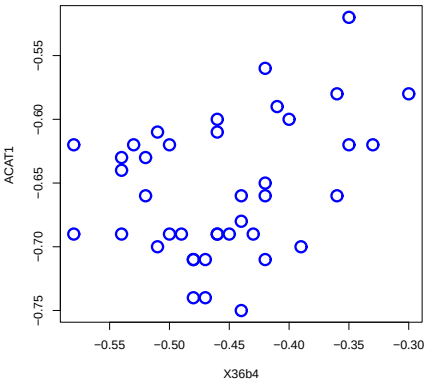

FIGURE 1 – *Souris : Nuage de points illustrant la faible liaison linéaire entre les expressions de deux gènes (corrélation de 0,33).*

Le choix des échelles à retenir pour réaliser un nuage de points peut s'avérer délicat. D'une façon générale, on distinguera le cas de variables *homogènes* (représentant la même grandeur et exprimées dans la même unité) de celui des variables *hétérogènes*. Dans le premier cas, on choisira la même échelle sur les deux axes (qui seront donc orthonormés) ; dans le second cas, il est recommandé soit de représenter les variables centrées et réduites sur des axes orthonormés, soit de choisir des échelles telles que ce soit sensiblement ces variables là que l'on représente (c'est en général cette seconde solution qu'utilisent, de façon automatique, les logiciels statistiques).

2.2 Rappel : variables centrées et réduites

Si X est une variable quantitative de moyenne $\overline{x}$ et d'écart–type σ_X, on appelle variable centrée associée à X la variable $X - \overline{x}$ (elle est de moyenne

nulle et d'écart–type σ_X), et variable centrée et réduite (ou tout simplement variable réduite) associée à X la variable $\frac{X-\overline{x}}{\sigma_X}$ (elle est de moyenne nulle et d'écart–type égal à un). Une variable centrée et réduite s'exprime sans unité.

2.3 Indice de liaison

Le coefficient de corrélation linéaire est un indice rendant compte numériquement de la manière dont les deux variables considérées varient simultanément. Il est défini à partir de la covariance qui généralise à deux variables la notion de variance :

$$\begin{aligned} \text{cov}(X,Y) &= \sum_{i=1}^{n} w_i [x_i - \overline{x}][y_i - \overline{y}] \\ &= \sum_{i=1}^{n} w_i x_i y_i - \overline{x}\,\overline{y}. \end{aligned}$$

La covariance est une forme bilinéaire symétrique qui peut prendre toute valeur réelle et dont la variance est la forme quadratique associée. En particulier, on en déduit les deux formules suivantes :

$$var(X+Y) = var(X) + var(Y) + 2cov(X,Y),$$

$$[cov(X,Y)]^2 \leq var(X)var(Y)\,;$$

(cette dernière propriété est l'inégalité de Cauchy-Schwarz).

la covariance dépend des unités de mesure dans lesquelles sont exprimées les variables considérées ; en ce sens, ce n'est pas un indice de liaison "intrinsèque".

C'est la raison pour laquelle on définit le coefficient de corrélation linéaire (appelé coefficient de Pearson ou de Bravais-Pearson), rapport entre la covariance et le produit des écarts-types :

$$\text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}.$$

Le coefficient de corrélation est égal à la covariance des variables centrées et réduites respectivement associées à X et Y : $\text{corr}(X,Y) = \text{cov}(\frac{X-\overline{x}}{\sigma_X}, \frac{Y-\overline{y}}{\sigma_Y})$.

Par conséquent, $\text{corr}(X,Y)$ est indépendant des unités de mesure de X et de Y. Le coefficient de corrélation est *symétrique* et prend ses valeurs entre -1 et +1. Les valeurs -1 et $+1$ correspondent à une liaison linéaire parfaite entre X et Y (existence de réels a, b et c tels que : $aX + bY + c = 0$).

Notons pour mémoire la possibilité d'utiliser d'autres indicateurs de liaison entre variables quantitatives. Construits sur les rangs (corrélation de Spearman) ils sont plus robustes faces à des situations de non linéarité ou des valeurs atypiques mais restent très réducteurs.

3 Une variable quantitative et une qualitative

3.1 Notations

Soit X la variable qualitative considérée, supposée à m modalités notées

$$x_1, \ldots, x_\ell, \ldots, x_m$$

et soit Y la variable quantitative de moyenne $\overline{y}$ et de variance σ_Y^2. Désignant par Ω l'échantillon considéré, chaque modalité x_ℓ de X définit une sous-population (un sous-ensemble) Ω_ℓ de Ω : c'est l'ensemble des individus, supposés pour simplifier de poids $w_i = 1/n$ et sur lesquels on a observé x_ℓ ; on obtient ainsi une *partition* de Ω en m classes dont nous noterons $n_1, \ldots, n_m$ les cardinaux (avec toujours $\sum_{\ell=1}^{m} n_\ell = n$, où $n = card(\Omega)$).

Considérant alors la restriction de Y à Ω_ℓ ($l = 1, \ldots, m$), on peut définir la moyenne et la variance partielles de Y sur cette sous-population ; nous les noterons respectivement $\overline{y_\ell}$ et σ_ℓ^2 :

$$\overline{y_\ell} = \frac{1}{n_\ell} \sum_{\omega_i \in \Omega_\ell} Y(\omega_i)\,;$$

$$\sigma_\ell^2 = \frac{1}{n_\ell} \sum_{\omega_i \in \Omega_\ell} [Y(\omega_i) - \overline{y_\ell}]^2.$$

3.2 Boîtes parallèles

Une façon commode de représenter les données dans le cas de l'étude simultanée d'une variable quantitative et d'une variable qualitative consiste à réaliser

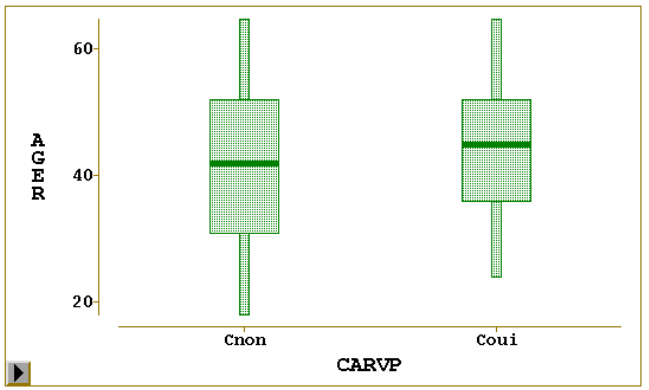

FIGURE 2 – *Banque : Diagrammes-boites illustrant les différences de distribution des âges en fonction de la possession d'une carte Visa Premier.*

des diagrammes-boîtes parallèles ; il s'agit, sur un même graphique doté d'une échelle unique, de représenter pour Y un diagramme-boîte pour chacune des sous-populations définies par X. La comparaison de ces boîtes donne une idée assez claire de l'influence de X sur les valeurs de Y, c'est-à-dire de la liaison entre les deux variables.

3.3 Formules de décomposition

Ces formules indiquent comment se décomposent la moyenne et la variance de Y sur la partition définie par X (c'est-à-dire comment s'écrivent ces caractéristiques en fonction de leurs valeurs partielles) ; elles sont nécessaires pour définir un indice de liaison entre les deux variables.

$$\overline{y} = \frac{1}{n}\sum_{\ell=1}^{m} n_\ell \overline{y_\ell}\,;$$

$$\sigma_Y^2 = \frac{1}{n}\sum_{\ell=1}^{m} n_\ell(\overline{y_\ell}-\overline{y})^2 + \frac{1}{n}\sum_{\ell=1}^{m} n_\ell \sigma_\ell^2 = \sigma_E^2 + \sigma_R^2\,.$$

Le premier terme de la décomposition de σ_Y^2, noté σ_E^2, est appelé *variance expliquée* (par la partition, c'est-à-dire par X) ou *variance inter* (between) ; le second terme, noté σ_R^2, est appelé *variance résiduelle* ou *variance intra* (within).

3.4 Rapport de corrélation

Il s'agit d'un indice de liaison entre les deux variables X et Y qui est défini par :

$$s_{Y/X} = \sqrt{\frac{\sigma_E^2}{\sigma_Y^2}}\,;$$

X et Y n'étant pas de même nature, $s_{Y/X}$ n'est pas symétrique et vérifie $0 \leq s_{Y/X} \leq 1$. Cet encadrement découle directement de la formule de décomposition de la variance. Les valeurs 0 et 1 ont une signification particulière intéressante.

4 Deux variables qualitatives

4.1 Notations

On considère dans ce paragraphe deux variables qualitatives observées simultanément sur n individus. On suppose que la première, notée X, possède r modalités notées $x_1, \ldots, x_\ell, \ldots, x_r$, et que la seconde, notée Y, possède c modalités notées $y_1, \ldots, y_h, \ldots, y_c$.

Ces données sont présentées dans un tableau à double entrée, appelé *table de contingence*, dans lequel on dispose les modalités de X en lignes et celles de Y en colonnes. Ce tableau est donc de dimension $r \times c$ et a pour élément générique le nombre $n_{\ell h}$ d'observations conjointes des modalités x_ℓ de X et y_h de Y ; les quantités $n_{\ell h}$ sont appelées les *effectifs conjoints*.

Une table de contingence se présente donc sous la forme suivante :

	y_1	$\cdots$	y_h	$\cdots$	y_c	sommes
x_1	n_{11}	$\cdots$	n_{1h}	$\cdots$	n_{1c}	n_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_ℓ	$n_{\ell 1}$	$\cdots$	$n_{\ell h}$	$\cdots$	$n_{\ell c}$	$n_{\ell +}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_r	n_{r1}	$\cdots$	n_{rh}	$\cdots$	n_{rc}	n_{r+}
sommes	n_{+1}	$\cdots$	n_{+h}	$\cdots$	n_{+c}	n

Les quantités $n_{\ell +} (\ell = 1, \ldots, r)$ et $n_{+h} (h = 1, \ldots, c)$ sont appelées les *effectifs marginaux* ; ils sont définis par $n_{\ell +} = \sum_{h=1}^{c} n_{\ell h}$ et $n_{+h} = \sum_{\ell=1}^{r} n_{\ell h}$, et ils vérifient $\sum_{\ell=1}^{r} n_{\ell +} = \sum_{h=1}^{c} n_{+h} = n$. De façon analogue, on peut définir les notions de fréquences conjointes et de fréquences marginales.

4.2 Représentations graphiques des profils

On peut envisager, dans le cas de l'étude simultanée de deux variables qualitatives, d'*adapter* les graphiques présentés dans le cas unidimensionnel : on découpe chaque partie (colonne, partie de barre ou secteur) représentant une modalité de l'une des variables selon les effectifs des modalités de l'autre. Mais, de façon générale, il est plus approprié de réaliser des graphiques représentant des quantités très utiles dans ce cas et que l'on appelle les *profils*.

On appelle ℓ-ème profil-ligne l'ensemble des fréquences de la variable Y conditionnelles à la modalité x_ℓ de X (c'est-à-dire définies au sein de la sous-population Ω_ℓ de Ω associée à cette modalité). Il s'agit donc des quantités :

$$\{\frac{n_{\ell 1}}{n_{\ell +}}, \ldots, \frac{n_{\ell h}}{n_{\ell +}}, \ldots, \frac{n_{\ell c}}{n_{\ell +}}\}.$$

On définit de façon analogue le h-ème profil-colonne :

$$\{\frac{n_{1h}}{n_{+h}}, \ldots, \frac{n_{\ell h}}{n_{+h}}, \ldots, \frac{n_{rh}}{n_{+h}}\}.$$

La représentation graphique des profils-lignes ou des profils-colonnes, au moyen, par exemple, de diagrammes en barre parallèles (*mosaïc plot*), donne alors une idée assez précise de la variation conjointe des deux variables.

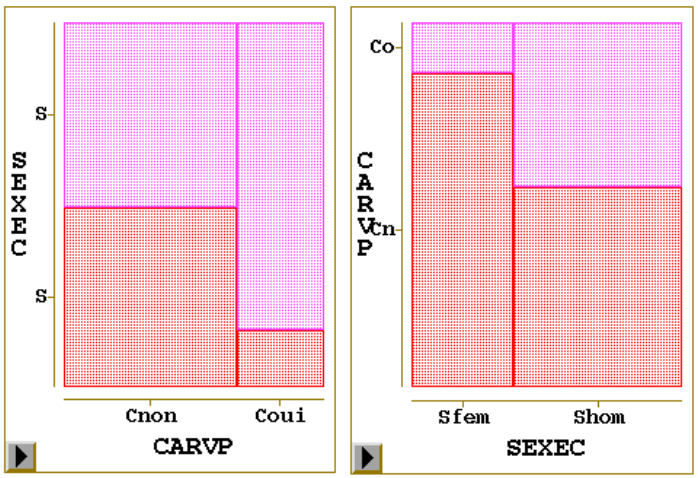

FIGURE 3 – *Banque : Diagrammes en barres des profils lignes et colonnes (mosaïque plot) de la table de contingence croisant le sexe et la possession de la carte Visa Premier. La superficie de chaque case est en plus proportionnelle à l'effectif de la cellule associée.*

4.3 Indices de liaison

Lorsque tous les profils-lignes sont égaux, ce qui est équivalent à ce que tous les profils-colonnes soient égaux et que

$$\forall (\ell, h) \in \{1, \ldots, r\} \times \{1, \ldots, c\} \ : n_{\ell h} = \frac{n_{\ell +} n_{+h}}{n} \ ,$$

on dit qu'il n'existe aucune forme de liaison entre les deux variables considérées X et Y. Par suite, la mesure de la liaison va se faire en évaluant l'écart entre la situation observée et l'état de non liaison défini ci-dessus.

4.3.1 Khi-deux

Il est courant en statistique de comparer une table de contingence observée, d'effectif conjoint générique $n_{\ell h}$, à une table de contingence donnée a priori (et appelée *standard*), d'effectif conjoint générique $s_{\ell h}$, en calculant la quantité

$$\sum_{\ell=1}^{r} \sum_{h=1}^{c} \frac{(n_{\ell h} - s_{\ell h})^2}{s_{\ell h}} \ .$$

De façon naturelle, pour mesurer la liaison sur une table de contingence, on utilise donc l'indice appelé khi-deux (chi-square) et défini comme suit :

$$\chi^2 = \sum_{\ell=1}^{r} \sum_{h=1}^{c} \frac{(n_{\ell h} - \frac{n_{\ell +} n_{+h}}{n})^2}{\frac{n_{\ell +} n_{+h}}{n}} = n \left[\sum_{\ell=1}^{r} \sum_{h=1}^{c} \frac{n_{\ell h}^2}{n_{\ell +} n_{+h}} - 1 \right] .$$

Le coefficient χ^2 est toujours positif ou nul et il est d'autant plus grand que la liaison entre les deux variables considérées est forte. Malheureusement, il dépend aussi des dimensions r et c de la table étudiée, ainsi que de la taille n de l'échantillon observé ; en particulier, il n'est pas majoré. C'est la raison pour laquelle on a défini d'autres indices, liés au khi-deux, et dont l'objectif est de palier ces défauts.

4.3.2 Autres indicateurs

Nous en citerons trois.

- Le *phi-deux* : $\Phi^2 = \frac{\chi^2}{n}$. Il ne dépend plus de n, mais dépend encore de r et de c.
- Le coefficient T de Tschuprow :

$$T = \sqrt{\frac{\Phi^2}{\sqrt{(r-1)(c-1)}}}.$$

  On peut vérifier : $0 \leq T \leq 1$.
- Le coefficient C de Cramer :

$$C = \sqrt{\frac{\Phi^2}{d-1}},$$

  avec : $d = \inf(r, c)$. On vérifie maintenant : $0 \leq T \leq C \leq 1$.

Enfin, la p-valeur d'un test d'indépendance (test du χ^2) est aussi utilisée pour comparer des liaisons entre variables.

5 Vers le cas multidimensionnel

L'objectif des prochains chapitres de ce cours est d'exposer les techniques de la statistique descriptive multidimensionnelle. Or, sans connaître ces techniques, il se trouve qu'il est possible de débuter une exploration de données multidimensionnelles en adaptant simplement les méthodes déjà étudiées.

5.1 Matrices des covariances et des corrélations

Lorsqu'on a observé simultanément plusieurs variables quantitatives (p variables, $p \geq 3$) sur le même échantillon, il est possible de calculer d'une part les variances de toutes ces variables, d'autre part les $\frac{p(p-1)}{2}$ covariances des variables prises deux à deux. L'ensemble de ces quantités peut alors être disposé dans une matrice carrée ($p \times p$) et symétrique, comportant les variances sur la diagonale et les covariances à l'extérieur de la diagonale ; cette matrice, appelée matrice des variances-covariances (ou encore matrice des covariances) sera notée $\mathbf{S}$. Elle sera utilisée par la suite, mais n'a pas d'interprétation concrète. Notons qu'il est possible de vérifier que $\mathbf{S}$ est semi définie positive.

De la même manière, on peut construire la matrice symétrique $p \times p$, comportant des 1 sur toute la diagonale et, en dehors de la diagonale, les coefficients de corrélation linéaire entre les variables prises deux à deux. Cette matrice est appelée matrice des corrélations, elle est également semi définie positive, et nous la noterons $\mathbf{R}$. Elle est de lecture commode et indique quelle est la structure de corrélation des variables étudiées.

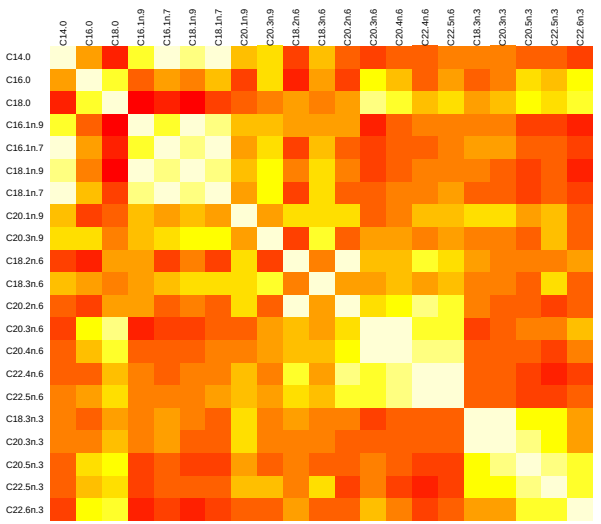

FIGURE 4 – *Souris : représentation graphique des corrélations entre les variables de concentration de lipides par des intensités de couleur.*

5.2 Tableaux de nuages

Notons $X^1, \ldots, X^p$ les p variables quantitatives considérées ; on appelle tableau de nuages le graphique obtenu en juxtaposant, dans une sorte de matrice carrée $p \times p$, p^2 sous-graphiques ; chacun des sous-graphiques diagonaux est relatif à l'une des p variables, et il peut s'agir, par exemple, d'un histogramme ; le sous-graphique figurant dans le bloc d'indice $(j, j'), j \neq j'$, est le nuage de points réalisé avec la variable X^j en abscisses et la variable $X^{j'}$ en ordonnées. Dans certains logiciels anglo-saxons, ces graphiques sont appelés *splom* (Scatter PLOt Matrix). Le tableau de nuages, avec la matrice des corrélations, fournit ainsi une vision globale des liaisons entre les variables étudiées.

5.3 La matrice des coefficients de Tschuprow (ou de Cramer)

Considérons maintenant le cas où l'on étudie simultanément plusieurs variables qualitatives (p variables, $p \geq 3$). La matrice des coefficients de Tschuprow est la matrice carrée d'ordre p, symétrique, comportant des 1 sur la diagonale et, en dehors de la diagonale, les coefficients de Tschuprow entre les variables prises deux à deux. Il s'agit donc d'une matrice du même type que la matrice des corrélations (elle est d'ailleurs, elle aussi, semi définie positive), et son utilisation pratique est analogue. Notons que l'on peut, de la même façon, utiliser les coefficients de Cramer au lieu des coefficients de Tschuprow.

5.4 Le tableau de Burt

Le tableau de Burt est une généralisation particulière de la table de contingence dans le cas où l'on étudie simultanément p variables qualitatives. Notons $X^1, \ldots, X^p$ ces variables, appelons c_j le nombre de modalités de $X^j, j = 1, \ldots, p$ et posons $c = \sum_{j=1}^{p} c_j$. Le tableau de Burt est en fait une matrice carrée $c \times c$, constituée de p^2 sous–matrices. Chacune des p sous–matrices diagonales est relative à l'une des p variables ; la $j^{\text{ième}}$ d'entre elles est carrée d'ordre c_j, diagonale, et comporte sur la diagonale les effectifs marginaux de X^j. La sous–matrice figurant dans le bloc d'indice $(j, j'), j \neq j'$, est la table de contingence construite en mettant X^j en lignes et $X^{j'}$ en colonnes ; le tableau de Burt est donc symétrique. Il apparaît en fait comme l'analogue qualitatif du tableau des nuages.