Séance 3: Liaisons entre variables

Analyse des individus

Sébastien Gadat

Laboratoire de Statistique et Probabilités
UMR 5583 CNRS-UPS

www.lsp.ups-tlse.fr/gadat

Troisième partie III

Liaisons entre variables

Objectifs

- Contexte : liaison entre variables numériques ou non
- Soit on ne dispose que de relations d'ordre entre les modalités de ces variables. Exemple : réponses à un sondage $\{A, B, C, D, E\}$
- Soit la valeur numérique d'une variable a peu d'importance, seul l'ordre importe. Exemple : avoir 12 en français ne signifie pas valoir deux fois plus que celui qui a 6
- Données : n individus décrits par 2 variables

$$X = \begin{pmatrix} 1 & 2 & \dots & n \\ r_1 & r_2 & \dots & r_n \\ s_1 & s_2 & \dots & s_n \end{pmatrix}$$

- Les r_i et s_i sont des permutations différentes des n entiers correspondant à chacun des individus
- Objectif : Étudier la dépendance entre les deux caractères décrivant X

Corrélation des rangs de Spearman

- On étudie la quantité

$$r_s = \frac{Cov(r,s)}{s_r s_s}$$

où s_r et s_s désignent les écarts types des quantités r et s.

- On définit

$$\forall i \in \{1 \dots n\} \qquad d_i = r_i - s_i$$

- Proposition : La quantité r_s est donnée par

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

Corrélation des rangs de Spearman

Interprétation :

- $r_s = 1$ si et seulement si les classements sont identiques.
- $r_s = -1$ si et seulement si les classements sont l'inverses l'un de l'autre.
- $r_s = 0$ si et seulement si les deux classements sont indépendants.
- Région critique de test de dépendance :
  - La région critique est de la forme $r_s > k$
  - k est obtenu via une table numérique
  - La table numérique contient en général les valeurs de r satisfaisant

$$P(r_s > r \quad | \quad \text{indépendance des variables}) = \alpha$$

    pour différentes valeurs de α
  - Exemple : pour $n = 50$, on mesure $r_s = 0.3$, or on lit que $P(r_s > 0.279 \quad | \quad \text{indépendance des variables}) = 0.05$. Conclusion ?
- Approximation pour n grand : on admettra que pour $n > 100$, on a $r_s \sim \mathcal{N}(0, \frac{1}{n-1})$

Corrélation des rangs τ de Kendall

- X et Y deux variables aléatoires décrites sur n individus.
- Signe du produit :

$$(X_i - X_j)(Y_i - Y_j)$$

où les individus i et j sont décrits par (X_i, Y_i) et (X_j, Y_j).

- Lorsque l'on effectue des réalisations indépendantes de (X, Y) notées (X_1, Y_1) et (X_1, Y_1), on définit :

$$\tau = 2P((X_1 - X_2)(Y_1 - Y_2) > 0) - 1$$

- $\tau \in [-1; 1]$ et s'annule lorsque X et Y sont des variables indépendantes (pourquoi ?)
- Idée : si τ est positif, X et Y ont plus de chances d'observer une variation dans le même sens.

Corrélation des rangs τ de Kendall

- Calcul du coefficient empirique sur n individus
- On considère tous les couples d'individus. Si les ordres concordent, on affecte $+1$ au couple, et sinon -1.
- Exemples :

$$I_1 : (-1, 3) \quad I_2 : (-10, 4) \longmapsto +1 \qquad I_1 : (-1, 3) \quad I_2 : (-10, 2) \longmapsto -1$$

- On additionne toutes ces valeurs pour tous les $n(n-1)/2$ couples distincts. Cette somme est notée S.
- $$S \in [-n(n-1)/2; n(n-1)/2]$$
- $$\hat{\tau} = \frac{2S}{n(n-1)}$$
- $\hat{\tau} = 1$: classements identiques. $\hat{\tau} = -1$: classements inversés
- Dès que $n \geq 8$, on considère que

$$\tau \sim \mathcal{N}\left((0; \frac{2(2n+5)}{9n(n-1)}\right)$$

Corrélation des rangs τ de Kendall : Méthode de calcul rapide

1. On ordonne les X_i par ordre croissants.
2. On compte le nombre de Y_j tels que $Y_j > Y_i$ avec $j > i$.
3. On additionne sur tous les i, et on pose R cette somme.
4. On a :

$$S = 2R - \frac{n(n-1)}{2} \qquad \hat{\tau} = \frac{4R}{n(n-1)} - 1$$

Exemple : Calculer les corrélations de rang r_s et $\hat{\tau}$ dans le cas suivant

X	1	2	3	4	5	6	7	8	9	10
Y	3	1	4	2	6	5	9	8	10	7

Au seuil 5%, les valeurs critiques sont $+/-0.648$ pour r_s et $+/-0.49$ pour τ, conclusion ?

Corrélation des rangs τ de Kendall : Méthode de calcul rapide

1. On ordonne les X_i par ordre croissants.
2. On compte le nombre de Y_j tels que $Y_j > Y_i$ avec $j > i$.
3. On additionne sur tous les i, et on pose R cette somme.
4. On a :

$$S = 2R - \frac{n(n-1)}{2} \qquad \hat{\tau} = \frac{4R}{n(n-1)} - 1$$

Exemple : Calculer les corrélations de rang r_s et $\hat{\tau}$ dans le cas suivant

X	1	2	3	4	5	6	7	8	9	10
Y	3	1	4	2	6	5	9	8	10	7

Au seuil 5%, les valeurs critiques sont $+/-0.648$ pour r_s et $+/-0.49$ pour τ, conclusion ?

Corrélation des rangs τ de Kendall : Méthode de calcul rapide

1. On ordonne les X_i par ordre croissants.
2. On compte le nombre de Y_j tels que $Y_j > Y_i$ avec $j > i$.
3. On additionne sur tous les i, et on pose R cette somme.
4. On a :

$$S = 2R - \frac{n(n-1)}{2} \qquad \hat{\tau} = \frac{4R}{n(n-1)} - 1$$

Exemple : Calculer les corrélations de rang r_s et $\hat{\tau}$ dans le cas suivant

X	1	2	3	4	5	6	7	8	9	10
Y	3	1	4	2	6	5	9	8	10	7

Au seuil 5%, les valeurs critiques sont $+/-0.648$ pour r_s et $+/-0.49$ pour τ, conclusion ?

Corrélation des rangs τ de Kendall : Méthode de calcul rapide

1. On ordonne les X_i par ordre croissants.
2. On compte le nombre de Y_j tels que $Y_j > Y_i$ avec $j > i$.
3. On additionne sur tous les i, et on pose R cette somme.
4. On a :

$$S = 2R - \frac{n(n-1)}{2} \qquad \hat{\tau} = \frac{4R}{n(n-1)} - 1$$

Exemple : Calculer les corrélations de rang r_s et $\hat{\tau}$ dans le cas suivant

X	1	2	3	4	5	6	7	8	9	10
Y	3	1	4	2	6	5	9	8	10	7

Au seuil 5%, les valeurs critiques sont $+/-0.648$ pour r_s et $+/-0.49$ pour τ, conclusion ?

Corrélation des rangs de Kendall, extension au cas de p variables

- On construit le tableau contenant les rangs pour chacun des individus sur les p variables
- $$X = \begin{pmatrix} r_{1,1} & r_{1,2} & \dots & r_{1,p} \\ r_{2,1} & r_{2,2} & \dots & r_{2,p} \\ \vdots & \dots & & \vdots \\ r_{n,1} & \dots & & r_{n,p} \end{pmatrix}$$
- Chaque colonne du tableau a pour somme $n(n+1)/2$
- La somme totale des éléments du tableau est $r_{.,.} = pn(n+1)/2$
- Si les classements étaient identiques, on aurait des sommes par ligne égales à $p, 2p, \dots, np$ à une permutation près.
- Si les classements étaient indépendants, on aurait des sommes par colonne identiques
- On définit : $S = \sum_{i=1}^{n} \left(r_{.,i} - \frac{r_{.,.}}{n} \right)^2$

Corrélation des rangs de Kendall, extension au cas de p variables

On vérifie que S est maximal s'il y a concordance des classements et dans ce cas :

$$S_{max} = \frac{p^2(n^3 - n)}{12}$$

Le coefficient de concordance des rangs se définit par

$$W = \frac{12S}{p^2(n^3 - n)}$$

La statistique W vérifie :

$$W \in [0; 1]$$

W faible correspond à l'indépendance des classements.
Si r_s désigne les coefficients de rangs de Spearman, on obtient la relation

$$\bar{r}_s = \frac{pW - 1}{p - 1}$$

où $\bar{r}_s$ est la moyenne des C_p^2 coefficients de Spearman.

Liaison entre variable numérique et variable qualitative

- Y quantitative et $\mathcal{X}$ qualitative
- Rapport de corrélation défini par :

$$\eta_{Y/\mathcal{X}} = \frac{V(E(Y|\mathcal{X}))}{V(Y)}$$

- Rapport de corrélation empirique, on énumère les modalités :

$$\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, \ldots \mathcal{X}_k\}$$

- Effectifs n_i de chaque modalités $\mathcal{X}_i$
- Moyenne $\bar{y}_i$ de Y sur les individus ayant la modalité $\mathcal{X}_i$
- Moyenne totale $\bar{y}$ de Y
-

$$e^2 = \frac{\frac{1}{n}\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2}{s_y^2}$$

Liaison entre variable numérique et variable qualitative

- $e^2 = 0$ si $\bar{y_1} = \bar{y_2} = \cdots = \bar{y_k} = \bar{y}$ d'où absence de dépendance
- $e^2 = 1$ si Y constant sur chacune des modalités (pourquoi ?)
- Ainsi, plus e^2 grand, plus il y a dépendance
- Cas à 2 classes :

$$e^2 = \frac{\frac{n_1 n_2}{n}(\bar{y_1} - \bar{y_2})^2}{s_y^2}$$

Liaison entre deux variables qualitatives

- X et Y deux variables qualitatives à r et s modalités
- On présente les données sous la forme d'une table de contingence :

	y_1	$\cdots$	y_h	$\cdots$	y_s	sommes
x_1	n_{11}	$\cdots$	n_{1h}	$\cdots$	n_{1s}	n_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_ℓ	$n_{\ell 1}$	$\cdots$	$n_{\ell h}$	$\cdots$	$n_{\ell s}$	$n_{\ell+}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_{m_1}	n_{r1}	$\cdots$	n_{rh}	$\cdots$	n_{rs}	n_{r+}
sommes	n_{+1}	$\cdots$	n_{+h}	$\cdots$	n_{+s}	n

- Les n_{i+} et n_{+j} s'appellent respectivement marges en lignes et colonnes.

Liaison entre deux variables qualitatives

On appelle profils-lignes le tableau des fréquences conditionnelles $\frac{n_{i,j}}{n_{i,+}}$. On a une définition équivalente pour les profils colonnes.
On remarque les propriétés importantes :

$$\sum_{i=1}^{r} \frac{n_{i,j}}{n_{i,+}} \left(\frac{n_{i,+}}{n}\right) = \frac{n_{+,j}}{n}$$

et

$$\sum_{i=1}^{r} \frac{n_{i,j}}{n_{+,j}} \left(\frac{n_{+,j}}{n}\right) = \frac{n_{i,+}}{n}$$

Autrement dit : les moyennes des profils lignes avec des poids correspondant aux effectifs marginaux des lignes correspond au profil marginal des colonnes et réciproquement.

Écart à l'indépendance

- Lorsque tous les profils lignes sont identiques, on parle d'indépendance entre X et Y.
- En effet, dans ce cas, les distributions conditionnelles sachant X ne sont pas modifiées pour Y.
- Cela se traduit par :

$$\frac{n_{1,j}}{n_{i,+}} = \frac{n_{2,j}}{n_{2,+}} = \cdots = \frac{n_{r,j}}{n_{r,+}}$$

- Dans le cas d'une indépendance entre X et Y, on peut d'ailleurs écrire :

$$n_{i,j} = \frac{n_{i,+} n_{+,j}}{n}$$

- On mesure l'indépendance des caractères X et Y en évaluant l'écart entre $n_{i,j}$ et $\frac{n_{i,+} n_{+,j}}{n}$.

Le χ^2 d'écart à l'indépendance

- On adopte la définition suivante :

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{i,j} - \frac{n_{i,+}n_{+,j}}{n}\right)^2}{\frac{n_{i,+}n_{+,j}}{n}}$$

- Proposition :

$$\chi^2 = n\left[\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{i,j}^2}{n_{i,+}n_{+,j}} - 1\right]$$

- Proposition :

$$\frac{\chi^2}{n} \leq \inf\{s-1; r-1\}$$

- Proposition :On a égalité dans l'inégalité précédente si et seulement si Y est fonctionnellement lié à X ou Y lié fonctionnellement à X.

Cas des tableaux 2×2

- Exercice : Donner la formule exacte dans le cas d'un tableau de taille 2×2 noté

$$\begin{array}{|cc|}\hline a & b \\ c & d \\ \hline\end{array}$$

- Dans le cas de variables X et Y indépendantes, on connaît la loi de la variable χ^2. Ainsi, on peut "tabuler" les quantités :

$$P\left(\chi^2 \geq M | H_0\right)$$

où H_0 désigne l'hypothèse d'indépendance des caractères X et Y.

- Cette tabulation permet d'effectuer des tests statistiques d'écart à l'indépendance appelés test du χ^2.