Licence d'Informatique 2
Analyse de Données Utilisateur (C5–160412)
TD 2 – Analyse Bidimensionnelle

Carl FRÉLICOT – Dpt Info / Lab MIA

L'objectif de l'analyse bidimensionnelle est d'étudier simultanément deux colonnes d'un tableau de données, de sorte de mettre en évidence la *forme* d'une éventuelle *liaison* (ou *dépendance*) entre les deux variables observées, et d'en évaluer l'*intensité*. On notera n le nombre d'observations (lignes du tableau), X et Y les deux variables observées.

La notion de *régression* sur des variables quantitatives renvoit à celle plus générale de modélisation à des fins de prévision. Dans la version la plus élémentaire, elle consiste à expliquer Y par une fonction de X[1] ; on parle de variable *expliquée* (*target*) et de variable *explicative* (*feature*) ou *prédicteur*. Si la fonction est affine, on parle de régression *linéaire*, et de régression *multiple* s'il s'agit d'une fonction à plusieurs variables explicatives. Il s'agit donc d'un problème d'*estimation* des *coefficients* du modèle à partir des observations. Dans le cas affine, on pose le modèle $\widehat{Y} = a\,X + b$, et la solution est donnée par : $\widehat{a} = \frac{s^2_{XY}}{s^2_X}$ et $\widehat{b} = \overline{y} - \widehat{a} \times \overline{x}$. Les valeurs $e_i = y_i - \widehat{y}_i$ sont appelés *résidus*.

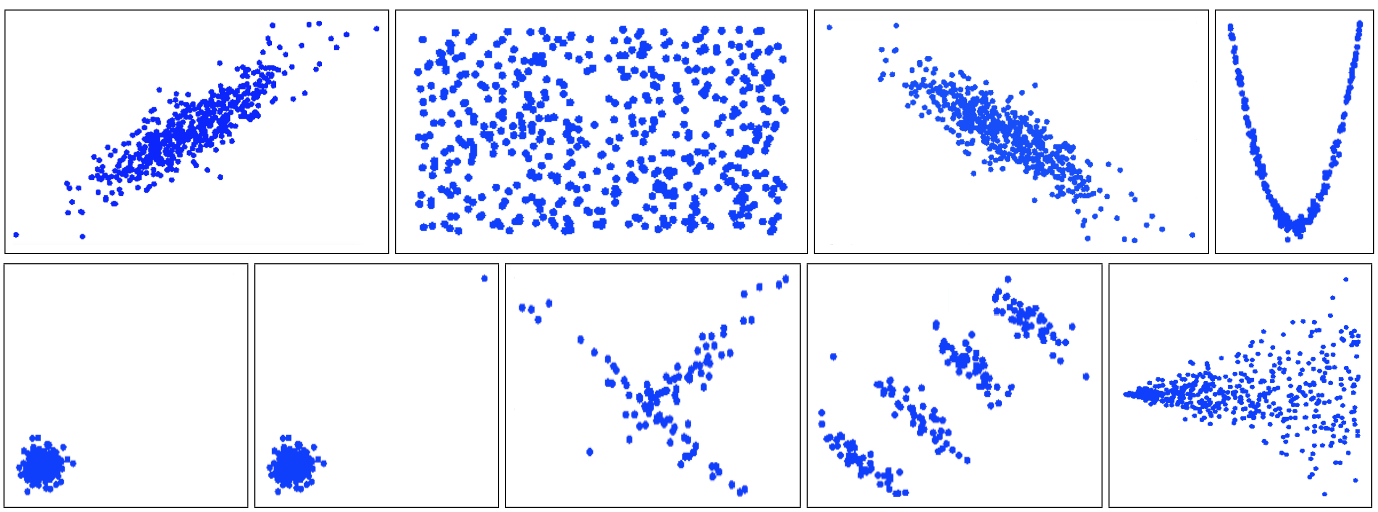
L'*analyse des correspondances* concerne les variables qualitatives. Elle s'attache à mettre en évidence les proximités et oppositions entre les modalités de l'une et celles de l'autre. On parle d'*analyse des correspondances multiples* dans le cas de plus de deux variables.

1. Deux variables quantitatives $(x_i, y_i)_{i=1,n}$

(a) examen graphique du nuage de points (*scatterplot*)
(b) connaissez-vous un/des indicateur/s numérique/s ?

Exercice 1

Pour chacun des nuages de points ci-dessous, qualifiez l'éventuelle liaison et estimez l'ordre de grandeur du coefficient de corrélation linéaire :



Exercice 2

Le tableau ci-dessous donne les distances d (en m) qu'il a fallu à une automobile roulant à diverses vitesses v (en km/h) pour s'arrêter. Le nuage de points (v_i, d_i) suggère d'utiliser un modèle quadratique de la vitesse pour expliquer la distance d'arrêt. On utilise donc $x = v^2$ pour expliquer $y = d$, de sorte que le modèle affine $y = a\,x + b$ est : $d = a\,v^2 + b$.

						$\Sigma\,.$	$\Sigma\,.^2$
v (km/h)	15	38	78	109	130	370	36 534
d (m)	3	15	60	112	180	370	48 778
$x = v^2$	225	1 444	6 084	1 1881	16 900	36 534	465 918 978
$x{\times}y = v^2{\times}d$	675	21 660	365 040	1 330 672	3 042 000	4 760 047	—

HW) Vous vérifierez chez vous que : $\overline{x} = 7\,306.8$, $s^2_x = 39\,794\,469.36$, et $\overline{y} = 74$.
2-1) Que vaut le coefficient de corrélation linéaire entre $y = d$ et $x = v^2$?
2-2) Estimez les coefficients du modèle.
2-3) Quelle distance d'arrêt faut-il prévoir si l'automobile a une vitesse de 150 km/h ?

[1] ou une transformation quelconque de X, par ex. : $\frac{1}{X}$, $\ln(X)$, X^3, $cos^2(X)$, etc

2-4) À quelle vitesse peut correspondre une distance d'arrêt de 90 m ?

2-5) Le manuel du code de la route préconise la méthode suivante pour calculer la distance d'arrêt : prendre le carré de la vitesse exprimée en dizaines de km/h. Qu'en pensez-vous ?

2-6) Que ce serait-il passé si une panne du capteur de vitesse (ou une erreur de saisie) avait transformé la valeur 130 en 1300 ? S'en serait-on aperçu et qu'aurait-on dû faire ?

2. Une variable quantitative et une variable qualitative $(x_i)_{i=1,n}$ et $(y_j)_{j=1,m}$

La variable Y définit une partition en m *groupes* des données. Chaque groupe correspond à une modalité de Y et est constitué de n_j $(j = 1, m)$ observations tel que $\sum_{j=1}^{m} n_j = n$.

(a) on peut calculer des indicateurs numériques sur X pour chacune des m modalités de Y, par ex. : moyennes $\overline{x}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_i$, et variances $s_j^2 = \frac{1}{n_j}\sum_{i=1}^{n_j}(x_i - \overline{x}_j)^2 = \frac{1}{n_j}\sum_{i=1}^{n_j} x_i^2 - \overline{x}_j^2$;
on montre que les statistiques globales (sans tenir compte de Y) se décomposent comme suit : moyenne globale $\overline{x} = \frac{1}{n}\sum_{j=1}^{m} n_j\,\overline{x}_j$, et variance globale $s^2 = \frac{1}{n}\sum_{j=1}^{m} n_j\,(\overline{x}_j - \overline{x})^2 + \frac{1}{n}\sum_{j=1}^{m} n_j\, s_j^2 = s_B^2 + s_W^2$, où
s_B^2 est la variance *inter-groupes*, c'est-à-dire la part de variance de X expliquée par Y, et s_W^2 est la variance *intra-groupes*, dite *résiduelle*, la part de variance de X non expliquée par Y.
À votre avis, quelle(s) caractéristique(s) possèdent ces indicateurs selon que X dépend de Y ou non ?
Peut-on définir une sorte de *"rapport de corrélation"* ayant des propriétés similaires (lesquelles ?) au coefficient de corrélation linéaire entre deux variables quantitatives, sauf qu'il n'est évidement pas symétrique ?

(b) Est-il possible de montrer graphiquement cette éventuelle dépendance ? Si oui, comment ?

Exercice 3

Des notes obtenues (X) selon trois méthodes d'enseignement (Y) sont données dans le tableau ci-dessous :

X	0	1.5	6	7.5	9	10	12	14	17	17	18	19
Y	tradi.	TEA	tradi.	TEA	TEA	tradi.	tradi	Renf.	Renf.	tradi.	Renf.	Renf.

Ci-contre, un autre tableau présentant les mêmes données, avec les statistiques dites *conditionnelles*, c'est-à-dire selon les modalités.

						n_j	$\overline{x}_j$	s_j^2
TEA	9	7.5	1.5			3	6	10.5
Tradi.	0	10	6	12	17	5	9	32.8
Renf.	14	17	19	18		4	17	3.5

3-1) Calculez les variances *inter* – et *intra-groupes*.

3-2) Déduisez le *"rapport de corrélation"* entre les notes et les méthodes d'enseignement.

HW) Vous calculerez ce qu'il faut pour tracer les *boxplot* parallèles, puis vérifierez sous Orange.

3. Deux variables qualitatives $(x_i)_{i=1,l}$ et $(y_j)_{j=1,c}$

Les variables X (à l modalités) et Y (à c modalités) sont observées sur n individus. Les données peuvent être présentées sous la forme d'un tableau à double entrée, appelé *table de contingence* et souvent noté N, qui croise les modalité de X (en ligne) et celles de Y (en colonne). Son terme général est le nombre d'individus n_{ij} pour lesquels on observe *conjointement* les modalité x_i et y_j. Les sommes en colonne $n_{i\bullet} = \sum_{j=1}^{c} n_{ij}$ et en ligne $n_{\bullet j} = \sum_{i=1}^{l} n_{ij}$ sont appelées *effectifs marginaux* respectivement de Y et X, et vérifient $\sum_{i=1}^{l} n_{i\bullet} = \sum_{j=1}^{c} n_{\bullet j} = n$. De manière analogue, on peut définir les *fréquences conjointes* et les *fréquences marginales*.

À partir du tableau N, on peut étudier les l profils-ligne $l_i = \left[l_{ij} = \frac{n_{ij}}{n_{i\bullet}}\right]_{j=1,c}$ et les c profils-colonne $c_j = \left[c_{ij} = \frac{n_{ij}}{n_{\bullet j}}\right]_{i=1,l}$ [2]

(a) connaissez-vous la propriété d'indépendance en calcul des probabilités ? Appliquez aux couples de modalités à partir de N.

(b) graphiquement, le tableau de données N fait correspondre deux nuages de points : en ligne, un nuage de l points en dimension c, et en colonne, un nuage de c points en dimension l ; on préfère s'intéresser aux nuages de profils. Les nuage de profils sont dans un espace de dimension $c-1$ et $l-1$; savez-vous pourquoi ?

(c) à votre avis, que valent les profils en cas d'indépendance parfaite ?

(d) connaissez-vous un/des indicateur/s numérique/s permettant de caractériser une dépendance entre deux variables qualitatives ?

Exercice 4

Considérons le résultat obtenu en jury (Z) à partir des notes de l'exercice précédent :

Z	ajourné	ajourné	ajourné	ajourné	reçu	reçu	reçu	mention	mention	mention	mention	mention

4-1) Déterminez la table de contingence $N = Y \times Z$.

4-2) Calculez le tableau des profils-ligne, puis dessinez le nuage correspondant.

4-3) Calculez et placez le profil moyen.

4-4) Faites les calculs permettant de juger si les résultats dépendent de la méthode ou non.

[2]dont les composantes sont respectivement les c fréquences des modalités y_j de Y conditionnellement à celles de $X = x_i$, et les l fréquences conditionnelles des modalité x_i de X conditionnellement à celles de $Y = y_j$