Analyse de données

Florent Bonneu

Chap 0 : Introduction

L'analyse de données c'est rigolo, ya des vieilles techniques pis des techniques un peu plus récentes.
Genre data-scientist c'est un métier récent et compliqué. Donc on va commencer par quelque chose d'un peu plus léger.

Principe :

Extraire des infos essentielles d'un énorme jeu de données.

Variables :

Variables quantitatives : mesurables (poids, age, revenus…).
- discrètes : nombre fini (nombre d'enfants, toussa toussa).
- continues : nombre infini ou très grand (revenus, toussa toussa).

Variables qualitatives : non mesurables (sexe, statut civil…).

Données :

- lignes : unités statistiques (objet, individu, pays).
- colonnes : variable observée par unité.

- ex (exploration) → ce qu'on va faire ici. Description statistique, analyse factorielle, classification. *Genre « la répartition semble quadratique (un U), exponentielle », toussa toussa.*
- ap (apprentissage) → modélisation statistiques pour la prédiction d'une variable.

Les trucs à faire :

- Extraction des données.
- Exploration des données.
- Exploration multi-dimensionnelle des données par des méthodes d'analyse factorielles et de classification.
- Partition aléatoire de l'échantillon.
- Estimation [blablabla]

Le but ici c'est d'extraire des données spécifiques d'une base de données et de voir si ya des corrélations (groupes d'individus avec des profils similaires).
Pour ce faire on va voir comment faire des représentations graphiques toutes jolies, on va aussi apprendre à faire de la réduction de variables (on passe de 100 variables à 4, c'est pratique et ça nous prend moins de huit heures à comprendre le schmilblick).

Assez-t-il tout étant, il ne faut pas aller trop vite et bien penser qu'aucune méthode n'est meilleure qu'une autre (*ça veut dire qu'on peut coder salement ?*).
Il ne faut également pas oublier d'interpréter les résultats. Le but c'est que ça veuille dire quelque chose.
Programmation en R (*via Rstudio c'est quand même plus joli ?*).

Méthodes d'analyses

Techniques factorielles (réduction & dimension) :

- ACP (analyse en composantes principales) → la plus importante, cool pour comprendre la technique.
- d'autres trucs (AFD, AFC, AFCM, MDS…).

Chapitre 1 : Statistique Descriptive Uni- et Bi-dimensionnelle

Là en gros on doit « explorer » les données pour se familiariser avec elles et trouver les trucs illogiques pis les enlever pour avoir un graphique tout beau après.

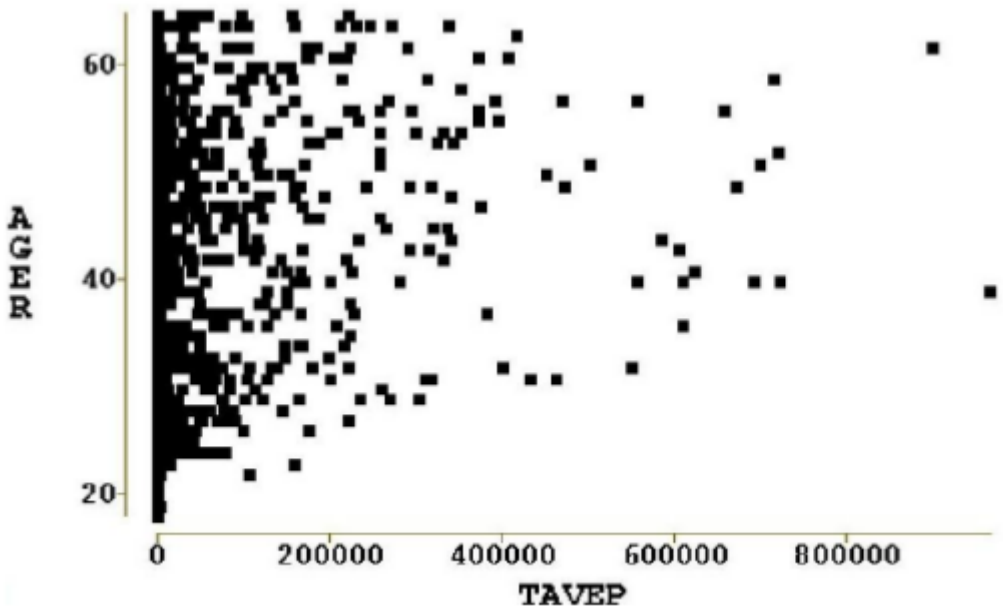
Pour ce faire la procédure comprend les représentations graphiques et les indicateurs numériques (on lit les données ou bien on les regarde).

- On peut s'amuser à regarder une seule variable avec des boites à moustache, des histogrammes et toutes ces sortes de choses vraiment géniales qui font qu'aujourd'hui vous êtes quelqu'un d'épanoui dans la vie.
- Et on peut regarder les liaisons entre variables qualitatives (coef de corrélation, ki2…).
- Il y a également des méthodes d'étude des relations entre variables de types différents (rapport de corrélation…).

Description de variables :

- Caractéristiques de position (*dont on va pas se soucier pasque le logiciel va tout gérer automatiquement*) : moyenne, médianne, quartile, mode (ou classe modale).
- Caractéristiques de dispersion (*idem*) : étendue, variance, écart-type, intervalles inter-quartile et inter-décile.
- Caractéristiques de symétrie et de normalité (*très peu utilisées ; concrètement on aura des graphiques qui permettront de voir ces trucs directement de visu*) : skewness (dissimétrie), kurtosis (applatissement), kolmogorov (tests de normalité).

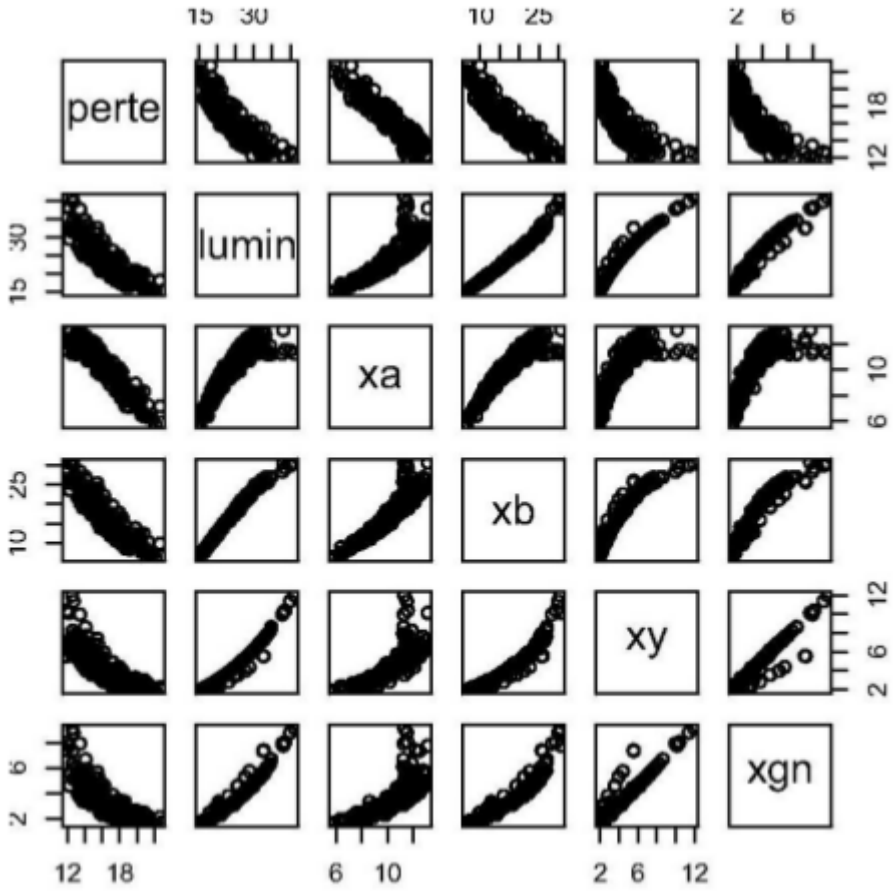
Ex :



Là on voit bien qu'il n'y a pas grand-chose à dire, mais qu'au moins ce n'est pas quadratique (le U) ni en forme exponentielle. Donc l'analyse va pas être trop compliquée normalement.

Premiers pas vers l'analyse multi-dimensionnelle

- Matrices de variance-covariance (ou des corrélations) : matrice symétrique $p \times p$ comprenant les coefficients de corrélation linéaire entre les variables prises deux à deux.
- Tableau de nuages : matrice carrée $p \times p$



Flemme d'expliquer du coup screenshot des diapos du cours.

Laisons entre variables

Entre variables quantitatives et qualitatives ;

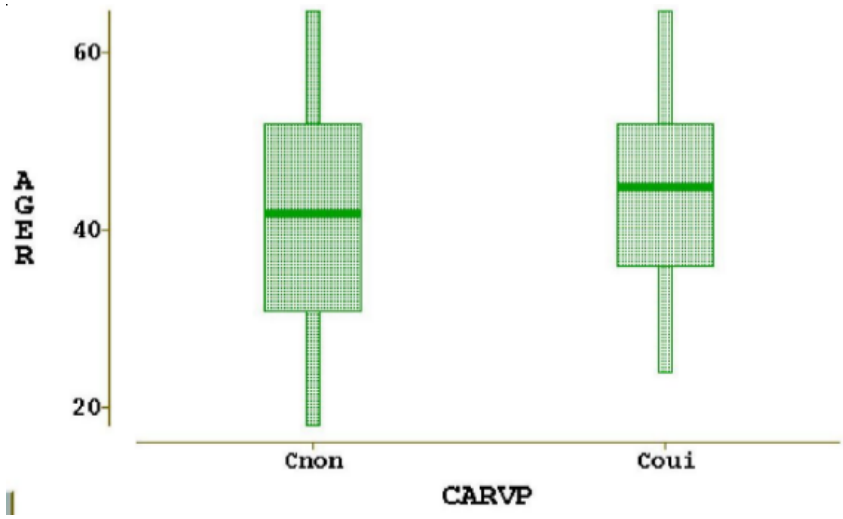
Notations

Soit X^i la variable qualitative supposée à r modalités notées $x_1^i, \cdots, x_r^i$.
Soit X^j la variable quantitative de moyenne $\bar{x}^j$ et de variance $\sigma^2_{x^j}$.

Principe de l'étude

L'ensemble des individus est découpé en r sous-populations à partir des r modalités de la variable X^i. Les caractéristiques de la variable quantitative X^j sont étudiées sur chaque population puis comparées.

On utilise ainsi les méthodes vues précédemment pour étudier les caractéristiques d'une variable quantitative sur chaque sous-population.



On cherche à trouver le(s) lien(s) entre les gros riches qui ont une CB visa première (VP), les autres, et leur âge.

Si vous voulez mon avis c'est compliqué.

Si les médianes sont au même niveau : **PAS DE LIEN** (pas d'influence de la modalité sur la distribution).

Si ya un décalage : **ELLES SONT LIÉES**.

Ici on voit qu'il y a un pitit décalage du coup on peut dire qu'elles sont **un peu** liées mais c'est pas flagrant.

Variance = **mesure de dispersion** (grande valeur = éloignement des données). Généralement on cherche la plus petite valeur pour les membres d'un groupe.

Variance **entre** les groupes = la plus grande possible (hisoire de bien séparer les gros pâtés).

Formule de décomposition

Illustre la décomposition de la moyenne et de la variance de X^j en fonction de celles calculées sur chacune des r sous-populations.

$$\begin{aligned} \bar{x}^j &= \frac{1}{n}\sum_{l=1}^{r} n_l \bar{x}_l^j \\ \sigma_{x^j}^2 &= \frac{1}{n}\sum_{l=1}^{r} n_l(\bar{x}_l^j - \bar{x}^j)^2 + \frac{1}{n}\sum_{l=1}^{r} n_l \sigma_l^2 = \sigma_E^2 + \sigma_R^2 \end{aligned}$$

où σ_E^2 est la variance expliquée par la partition (inter) et σ_R^2 est la variance résiduelle (intra).

Rapport de corrélation

$$s_{X^j/X^i} = \sqrt{\frac{\sigma_E^2}{\sigma_{X^j}^2}}$$

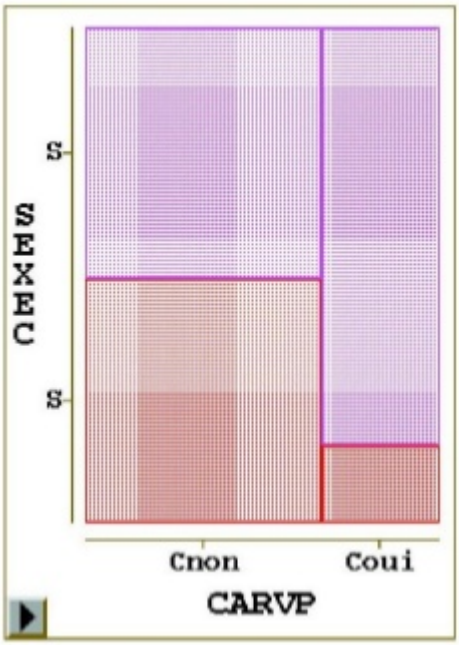
- $0 \leq s_{X^j/X^i} \leq 1$
- $s_{X^j/X^i} = 0$: pas de liaison et $s_{X^j/X^i} = 1$: liaison maximale.

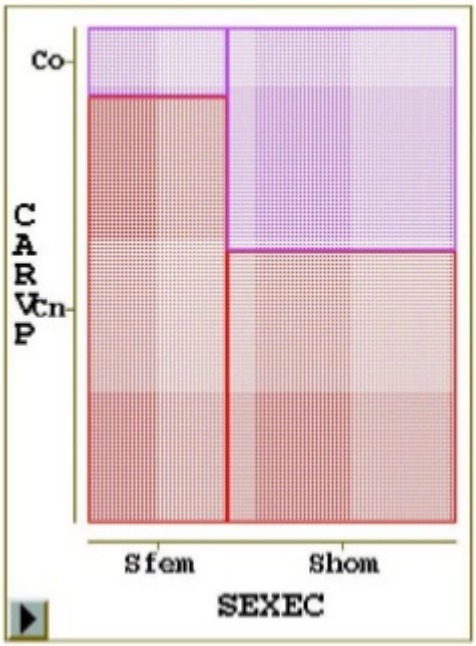
Table de contingence

Ce sera une table qui servira d'entrée à des scripts.

C'est une table qui compte les individus par rapport à des trucs.

Par exemple la proportion d'hommes qui ont une CB VP par rapport aux femmes (voir le truc en dessous).





On voit bien qu'en général ya plus d'hommes qui ont la carte (le truc rose sur Coui est vachement grand sur le graphe de gauche).

On peut lire ça comme ça :

« *Sur toutes les personnes qui **ont la carte**, **80** % sont des **hommes** (et **20** % sont des **femmes**), et sur toutes les personnes qui **ne l'ont pas**, **50** % sont des **hommes** et **50** % sont des **femmes**.* »

Par contre on voit que l'aire des boites ça représente le nombre de personnes. Donc via le graphe de droite on voit que les femmes sont un peu moins nombreuses à avoir été interrogées (la largeur est plus faible). On peut voir ça aussi sur le graphe de gauche mais c'est moins flagrant.

Indice de liaison

L'indice de Khi-deux est la somme des équarts quadratiques entre n_{lh} et [pas entendu].

- Aucune liaison entre X^i et X^j ssi :

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

Khi-deux : $$\chi^2 = \sum_{\ell=1}^{r} \sum_{h=1}^{c} \frac{\left(n_{\ell h} - \frac{n_{\ell+} n_{+h}}{n}\right)^2}{\frac{n_{\ell+} n_{+h}}{n}}$$

Phi-deux : $$\Phi^2 = \frac{\chi^2}{n}$$

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

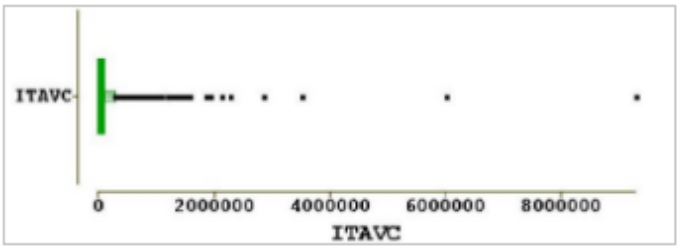
- On vérifie : $0 \leq T \leq 1$
- Plus la valeur de ces indices est grande et plus la liaison entre les deux variables considérées est forte.

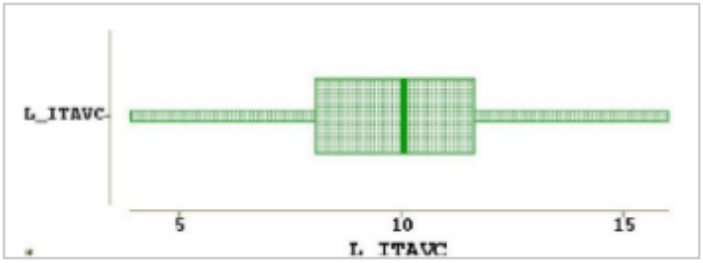
T = 0 → il y a abscence de liaison.

T = 1 → c'est très très lié.

On verra le test du Khi-deux aussi [voir cours de M1 de Cécile].

Transformation logarithmique





C’est rigolo ça permet de changer l’apparence des données affichées pour décrypter un peu mieux à l’œil nu.

Chapitre 2 :Analyse en Composantes Principales (ACP)

Voir exemple du cours (ci dessous).

Objectifs

Représenter graphiquement des résumés (nuages de points) de :

- $p > 3$ variables, en respectant la structure des corrélations entre variables,
- n individus, en respectant les distances entre individus.

Notes de $n = 9$ élèves dans $p = 4$ disciplines

	MATH	PHYS	FRAN	ANGL
jean	6.00	6.00	5.00	5.50
alan	8.00	8.00	8.00	8.00
anni	6.00	7.00	11.00	9.50
moni	14.50	14.50	15.50	15.00
didi	14.00	14.00	12.00	12.50
andr	11.00	10.00	5.50	7.00
pier	5.50	7.00	14.00	11.50
brig	13.00	12.50	8.50	9.50
evel	9.00	9.50	12.50	12.00

Statistiques élémentaires

Variable	Moyenne	Ecart-type	Minimum	Maximum
MATH	9.67	3.37	5.50	14.50
PHYS	9.83	2.99	6.00	14.50
FRAN	10.22	3.47	5.00	15.50
ANGL	10.06	2.81	5.50	15.00

Matrice des variances-covariances

	MATH	PHYS	FRAN	ANGL
MATH	11.39	9.92	2.66	4.82
PHYS	9.92	+ 8.94	4.12	5.48
FRAN	2.66	4.12	+ 12.06	9.29
ANGL	4.82	5.48	9.29	+ 7.91

Covariance(X,X) = Variance(X)

Coefficients de corrélation

	MATH	PHYS	FRAN	ANGL
MATH	1.00	0.98	0.23	0.51
PHYS	0.98	1.00	0.40	0.65
FRAN	0.23	0.40	1.00	0.95
ANGL	0.51	0.65	0.95	1.00

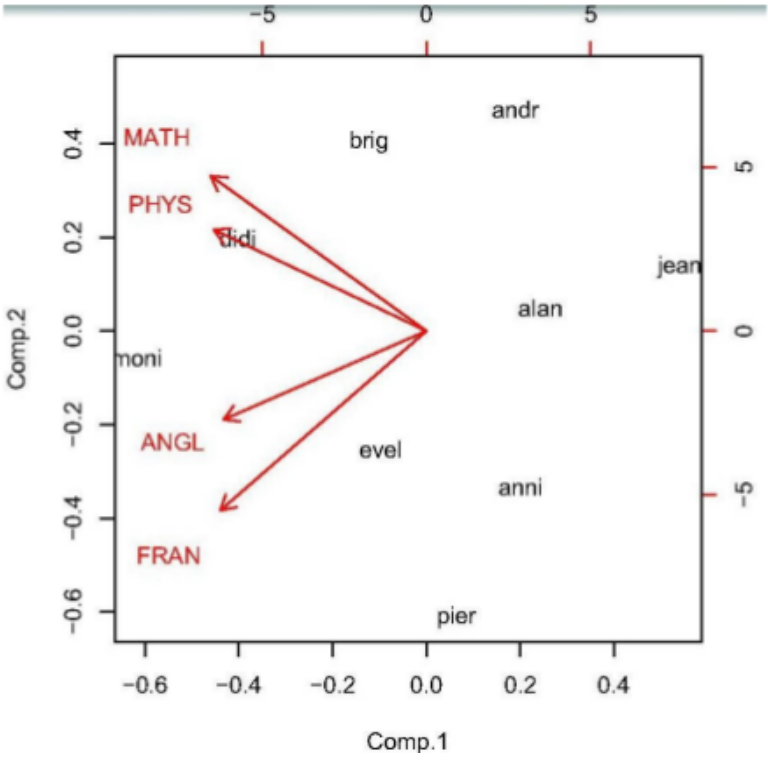
Les élèves qui ont eu des bonnes notes en maths n'ont pas vraiment eu de bonnes notes en anglais (0.51).

Les élèves qui ont eu de bonnes notes en anglais ont également eu de bonnes notes en Français (0.95).

Valeurs propres ; variances expliquées

FACTEUR	VAL. PR.	PCT. VAR.	PCT. CUM.
1	28.23	0.70	0.70
2	12.03	0.30	1.00
3	0.03	0.00	1.00
4	0.01	0.00	1.00

Voilà ce qu’on pourra avoir avec l’ACP :



Il y a eu réduction de dimension (on avait 4 variables et on passe à une représentation 2D (donc avec deux variables)).

Interprétation :

On voit que Jean a les pires notes de la classe (il est positionné à l’inverse des flèches).
Brigitte quand à elle a de meilleures notes dans les matières scientifiques que dans les matières littéraires.