Exploration des données

Antoine Nongaillard

But Info – IUT – Université de Lille

R5.AB.12 – Séance 04

Plan

Cycle de travail du data scientist

1. Récupération des données
   Les données peuvent être hétérogènes (image, son), de différents bases, voire nécessiter la création d'un vecteur de récupération.
2. Nettoyage des données (et regroupement, *data architect*)
   Les données doivent être consistantes, sans valeurs aberrantes ni manquantes, sous le même format, accessibles au même endroit et au bon moment.
3. Exploration des données (*data analyst*)
   Le but est de mieux comprendre les différents comportements et de bien saisir le phénomène sous-jacent.
4. Modélisation à partir des données (et utilisation d'algorithmes pour créer de l'intelligence (artificielle) qui aide à la décision.)
   Il convient de trouver un modèle (stochastique ou déterministe) du phénomène à l'origine des données.
5. Exploitation du modèle.

Plan

Qu'est-ce qu'une corrélation ?

En probabilités et en statistique, la corrélation entre plusieurs variables aléatoires ou statistiques est une notion de liaison qui contredit leur indépendance. Autrement dit : est-ce que les valeurs de X dépendent des valeurs de Y ? Ou est-ce que les valeurs de Y dépendent des valeurs de X ?

Dire que Y dépend de X signifie que la connaissance des valeurs de X permet de prédire, dans une certaine mesure, les valeurs de Y. En d'autres termes, si Y dépend de X, on peut trouver une fonction f telle que :

$$Y = f(X)$$

On dit que Y est la variable dépendante (à expliquer) et que X est la variable indépendante (explicative).

La notion de dépendance n'est pas symétrique !

Rechercher des corrélations

Pourquoi ?

Quelques exemples sur des résultats d'étudiants :

- Sachant qu'un individu a eu faux à la question 2, a-t-il de grandes chances d'avoir répondu faux, ou vrai, à la question 3 ?
- Étant donné les résultats obtenus en Mathématique au BAC, quelles chances un candidat a-t-il de réussir son premier semestre ?

Quelques exemples sur des opérations bancaires :

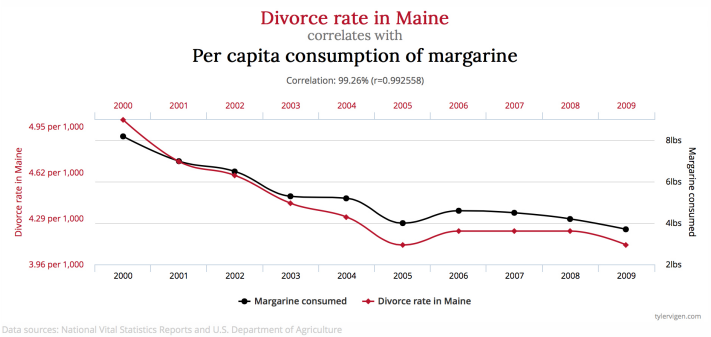
- avez-vous les mêmes catégories de dépenses le week-end et en semaine ?
- le montant d'une opération est-il différent d'une catégorie de dépense à l'autre ?
- y a-t-il des catégories d'opérations qui arrivent toujours au même moment du mois, comme votre loyer, par exemple ?
- vos paiements en carte bancaire sont-ils toujours petits, et vos virements importants ?

La nature des variables détermine la méthode de recherche des corrélations.

Corrélation et causalité

La corrélation entre deux variables correspond à la relation qu'il existe entre elles. Au niveau mathématique, cela revient à étudier la dépendance qu'il existerait entre les deux évènements ayant généré ces variables.

On peut avoir une corrélation sans avoir de lien de cause à effet [1]

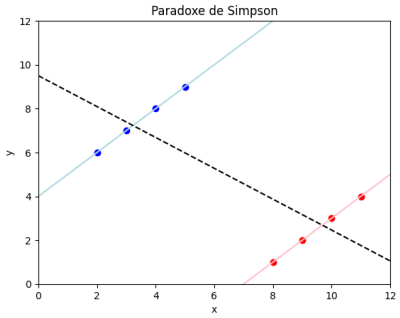


Célèbre citation de *Ronald Coase* : "*If you torture the data long enough, it will confess*".

1. D'autres exemples sur : http://www.tylervigen.com/spurious-correlations

Le paradoxe de Simpson

Aussi appelé *effet Yule-Simpson*, c'est un paradoxe statistique décrit par George Yule en 1903 et Edward Simpson en 1951, dans lequel un phénomène observé dans plusieurs groupes s'inverse lorsque les groupes sont combinés.



Le paradoxe de Simpson

Un exemple réel provenant d'une étude médicale sur le succès de deux traitements contre les calculs rénaux.

Succès du traitement selon la taille des calculs rénaux

Tâche	Traitement A	Traitement B
Petits calculs	93% (81/87)	87% (234/270)
Gros calculs	73% (192/263)	69% (55/80)

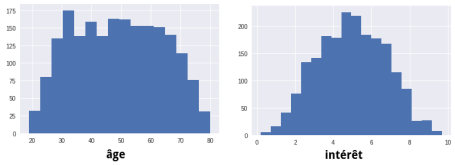
Traitement A	Traitement B
78% (273/350)	83% (289/350)

Le paradoxe vient du fait que le traitement A a été donné beaucoup plus souvent pour les gros calculs, qui sont plus difficiles à soigner.

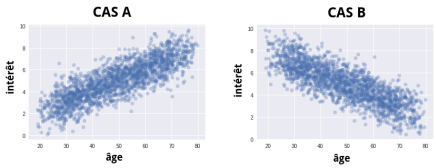
- la variable supplémentaire (ici la taille des calculs) a un impact significatif sur les rapports, elle a une influence en même temps sur le choix du traitement et sur le résultat du traitement
- les tailles des groupes combinés quand la variable supplémentaire est ignorée sont très différentes
- cette variable supplémentaire est appelé **facteur de confusion**

La distribution empirique est insuffisante !

Une analyse bi-variée permet d'établir la relation entre deux variables. On place l'une en abscisse, l'autre en ordonnée. Grâce à des diagrammes de dispersion (*scatter plot*), on peut voir apparaître des relations :



Deux cas sont envisageables à partir des mêmes distributions empiriques :



Le diagramme de corrélation

En amont de toute mesure, il est nécessaire de définir la forme d'une éventuelle relation à l'aide d'une représentation graphique appropriée. Selon la forme de la relation observée, on ne fera pas les mêmes hypothèses et on n'utilisera pas les mêmes outils de mesure.

Pour savoir s'il existe une relation entre deux caractères, on établit un **diagramme de corrélation**, croisant les modalités de X et de Y.

Le nuage des points de coordonnées (X_i, Y_i) permet de caractériser la relation à l'aide de trois critères :

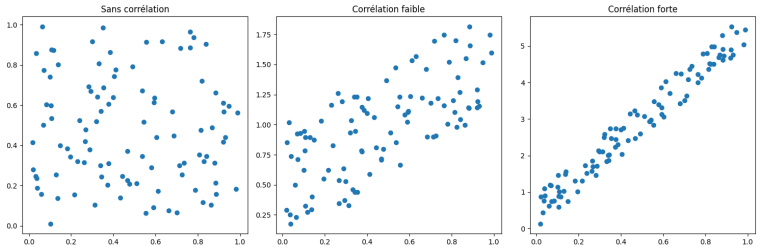
- l'intensité de la relation
- la forme de la relation
- le sens de la relation

Ces caractéristiques déterminent si le calcul d'un coefficient de corrélation est adéquat, et éventuellement selon quelle méthode...

Le diagramme de corrélation

L'intensité d'une relation

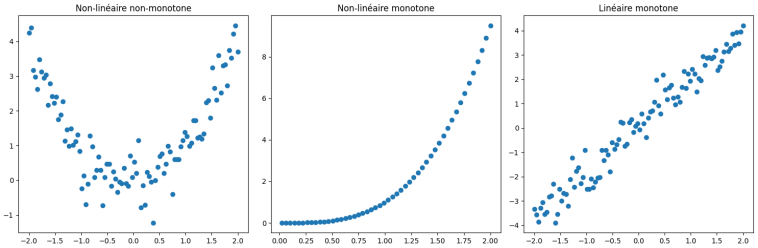
- **Une relation est forte** si les unités ayant des valeurs voisines sur X ont également des valeurs voisines sur Y, c'est à dire si X_i proche de $X_j \Rightarrow Y_i$ proche de Y_j
- **Une relation est faible** si les unités ayant des valeurs voisines sur X peuvent avoir des valeurs éloignées sur Y
- **Une relation est nulle** si les valeurs de X ne permettent aucunement de prédire les valeurs de Y.



Le diagramme de corrélation

La forme d'une relation

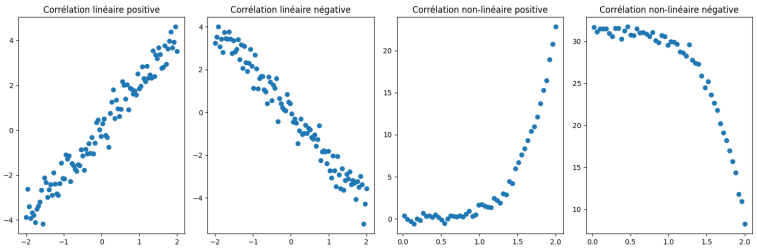
- **Une relation est linéaire** si l'on peut trouver une relation entre X et Y de la forme $Y = aX + b$, c'est à dire si le nuage de point peut s'ajuster correctement à une droite.
- **Une relation est non-linéaire** si la relation entre X et Y n'est pas de la forme $Y = aX + b$, mais de type différent (parabole, hyperbole, sinusoïde, etc). Le nuage de point présente alors une forme complexe avec des courbures.
- **Une relation est non-linéaire monotone** si elle est strictement croissante ou strictement décroissante, c'est-à-dire si elle ne comporte pas de minima ou de maxima. Toutes les relations linéaires sont monotones.



Le diagramme de corrélation

Le sens d'une relation

- **Une relation monotone (linéaire ou non) est positive** si les deux caractères varient dans le même sens, c'est à dire si $X_i > X_j \Rightarrow Y_i > Y_j$
- **Une relation monotone est négative** si les deux caractères varient en sens inverse, c'est à dire si $X_i > X_j \Rightarrow Y_i < Y_j$



Plan

Rechercher des corrélations

Entre variables quantitatives continues

Entre deux variables quantitatives, on utilise généralement un diagramme de dispersion. Si l'histogramme d'un sous-échantillon diffère beaucoup de celui général, alors il y a une corrélation !

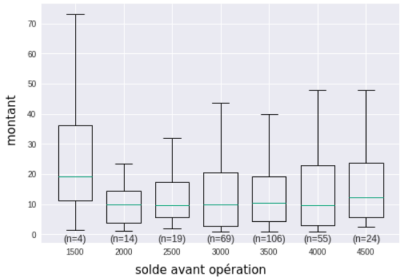
À gauche, pas de corrélation apparente. À droite, une corrélation évidente.

Rechercher des corrélations

Entre variables quantitatives continues

Il arrive que les points soient nombreux et dispersés : difficile d'y voir clair. Afin d'y remédier, il convient :

- d'agréger la variable X en différentes classes
- représenter ensuite chaque classe par un *boxplot*.



Attention à veiller aux effectifs ! Certaines boîtes sont assez dispersées mais ne portent que sur peu d'effectifs. Il y a un problème de **significativité**.

Rechercher des corrélations

Entre variables quantitatives continues

Un indicateur de base nécessaire est la **covariance empirique** de x et y.

$$S_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \overline{x}\cdot\overline{y}$$

La covariance s'interprète de la manière suivante :

- quand il y a corrélation entre x et y, $S_{x,y}$ sera proche de 0
- si x est petit quand y est grand (ou inversement), alors $S_{x,y}$ sera positif
- si x est grand quand y est petit (et inversement), alors $S_{x,y}$ sera négatif.

On normalise souvent un indicateur (ici, par le produit des écarts-types) souvent afin de faire des comparaisons.

Rechercher des corrélations

Entre variables quantitatives continues

Le **coefficient de corrélation linéaire** (ou de (*Bravais-*)*Pearson*), noté r, se définit par :

$$r_{x,y} = \frac{S_{x,y}}{S_x \cdot S_y}$$

Il prend des valeurs entre -1 et 1, son signe indique le sens de la relation alors que sa valeur absolue indique son intensité. Ce coefficient n'est applicable que pour mesurer la relation entre deux variables x et y ayant une distribution de type gaussien et ne comportant pas d'*outlier*.

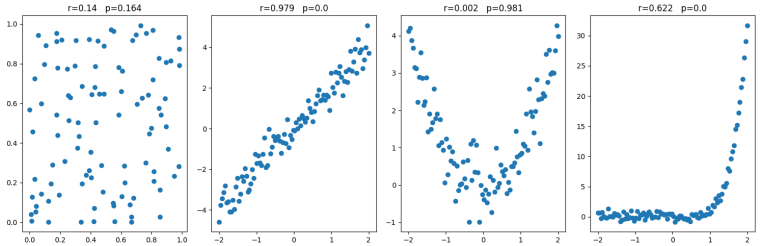
Le **coefficient de corrélation de rang** (ou de *Spearman*) examine s'il existe une relation à partir des rangs des observations. Il détecte les relations monotones quelle que soit leur forme.

$$\rho(X,Y) = 1 - \frac{6\cdot\left(\sum_{i=1}^{n}(rg(x_i) - rg(y_i))^2\right)}{n^3 - n}$$

Rechercher des corrélations

Entre variables quantitatives continues

Voici l'interprétation des indicateurs :



En python, les coefficients de corrélation linéaire et de rang s'obtiennent par :

```
import scipy.stats as st
import numpy as np

res1 = st.pearsonr(x1,y1)
r1,p1 = np.round(res1[0],3), np.round(res1[1],3)

alt1 = st.spearmanr(x1,y1)
R1,P1 = np.round(alt1[0],3), np.round(alt1[1],3)
```

En python...

Avec la librairie `pandas`, il est possible de calculer les coefficients de corrélation entre toutes les paires de variables numériques, grâce à `DataFrame.corr(...)`, ou d'une série avec d'autres grâce à `DataFrame.corrwith(...)` ::

```
df.corr(method='pearson', numeric_only=True)
```

	Age	Height	Weight	Year
Age	1.000000	0.137940	0.211718	-0.108380
Height	0.137940	1.000000	0.796213	0.047578
Weight	0.211718	0.796213	1.000000	0.019095
Year	-0.108380	0.047578	0.019095	1.000000

```
df_tmp = df.loc[:,['Height', 'Weight', 'Year']]
df_tmp.corrwith(df['Age'])
```

```
Height 0.137940
Weight 0.211718
Year -0.108380
dtype: float64
```


Significativité d'une corrélation

Un test de la **significativité**, notée *p*, d'une éventuelle relation et une vérification de la validité (absence de biais) est encore nécessaire.

Par exemple :

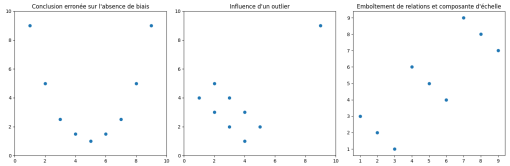
- un *r* de + 0.6 établi sur un échantillon de 10 observations n'est pas significatif au seuil de 5%
  - ⇒ il peut s'agir d'un hasard
- un *r* de + 0.2 établi sur un échantillon de 200 personnes est significatif au seuil de 5%
  - ⇒ la taille de l'échantillon fait que la relation, bien que faible a peu de chances d'être due au hasard

On réalise un test d'hypothèse :

1. `H` : "*il n'y a pas de relation entre X et Y*"
2. on fixe un risque d'erreur pour le rejet de `H` : α = 5%
3. on calcule $|r(x, y)|$
4. on calcule la valeur théorique $r(\alpha, n)$ qui n'est dépassé que dans α% des cas
5. on teste `H` vraie si $r(\alpha, n) > |r(x, y)|$
6. on accepte ou on rejette `H`

Significativité d'une corrélation

Le rejet d'une hypothèse d'indépendance ne doit pas amener à conclure trop vite à l'existence d'une relation. Elle peut souvent être la conséquence de biais liés à un mauvais respect des conditions d'utilisation des coefficients de corrélation.



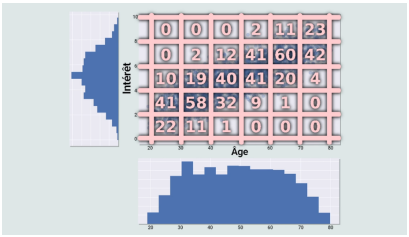
Quelques exemples :

- À gauche, un simple examen des coefficients sans avoir tracé le nuage de points ferait manquer la corrélation.
- Au centre, une corrélation (r = +0.54) mais non significative (au seuil de 5%), due uniquement à un outlier. Si on le retire, on obtient (r = $-$0.67) significative (au seuil de 5%).
- À droite une relation positive significative (r = +0.75), mais résultant de différents comportements de 3 sous-populations à l'intérieure desquelles la relation est rigoureusement négative.

Le tableau de contingence

Entre deux variables discrètes (quantitatives ou qualitatives), on établit un **tableau de contingence** pour déterminer la forme de la relation et on fait un test du *chi-2* pour sa significativité.

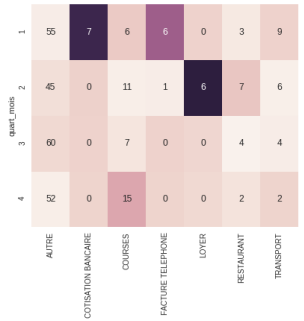
Pour obtenir un tableau de contingence, on découpe un diagramme de dispersion en cases où l'on dénombre les observations.

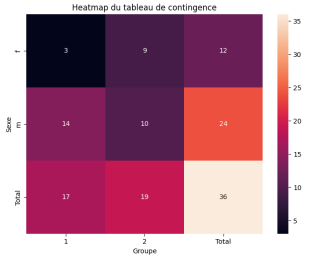


Rechercher des corrélations

Entre variables discrètes

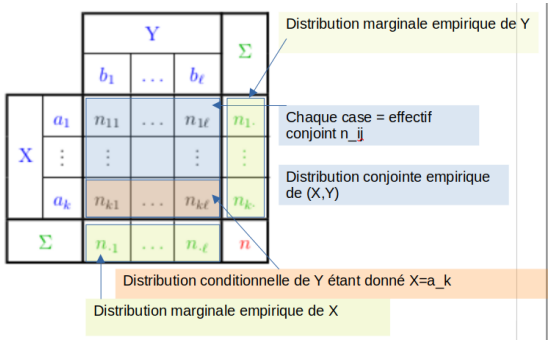
Les corrélations entre deux variables discrètes sont représentées par une "carte de chaleur" (*heatmap*).





En regardant ces cases foncées, on apprend que les cotisations bancaires et factures téléphoniques sont souvent payées en tout début de mois, que les loyers sont souvent payés en 2e quartier de mois. . .

Le tableau de contingence



- N_{ij}, effectif de la ligne i et de la colonne j
- $N_{i.}$, somme de la ligne i
- $N_{.j}$, somme de la ligne j
- $N_{..}$, somme générale du tableau (nombre d'observations)

Voici un exemple :

Sexe / Groupe	Groupe 1	Groupe 2	Total
Femme	3	9	12
Homme	14	10	24
Total	17	19	36

Indiquant des effectifs bruts, le tableau de contingence ne permet pas de comparer les proportions.

La proportion d'hommes est-elle plus élevée dans le groupe 1 que dans le groupe 2 ?

Le tableau de contingence

On construit donc généralement deux tableaux de profils indiquant les pourcentages en lignes ou les pourcentages en colonnes.

- en ligne : $N_{ij} \Rightarrow \frac{N_{ij}}{N_{i.}}$
- en colonne : $N_{ji} \Rightarrow \frac{N_{ij}}{N_{.j}}$

Sexe / Groupe	Groupe 1	Groupe 2	Total
Femme	3	9	12
Homme	14	10	24
Total	17	19	36

Sexe / Groupe	Groupe 1	Groupe 2	Total
Femme	18%	47%	33%
Homme	82%	53%	67%
Total	100%	100%	100%

Sexe / Groupe	Groupe 1	Groupe 2	Total
Femme	25%	75%	100%
Homme	58%	42%	100%
Total	47%	53%	100%

On remarquera qu'une même case du tableau de contingence peut toujours être décrite de deux façon différente. Si l'on prend la case N_{12} , elle indique que les 9 femmes du groupe 2 représentent 47% du groupe 2 et 75% des femmes de la population totale.

En python

Le tableau de contingence s'obtient par :

```
X='Genre'
Y='Groupe'
tab = df[[X,Y]].pivot_table(index=X,columns=Y,aggfunc=len,margins=True,margins_name="Total")
```

Les profils de lignes s'obtiennent simplement :

```
profiles_lig = tab.div(cont['Total'], axis=0)
```

Tout comme les profils de colonnes :

```
profiles_col = tab.div(tab.loc['Total'], axis=1)
```

Une *heatmap* s'obtient via `seaborn` :

```
sns.heatmap(tab, annot=True, fmt="d")
```

Rechercher des corrélations

Entre variables discrètes

On peut aussi comparer les effectifs observés de chacune des cases N_{ij} aux effectifs théoriques N^*_{ij} qui seraient obtenus s'il n'y avait aucun lien entre les deux modalités X et Y.

Calcul des effectifs théoriques : $N^*_{ij} = \frac{N_{i.} N_{.j}}{N_{..}}$

Afin de pouvoir décrire la forme d'une éventuelle relation entre les modalités de X et de Y, on peut calculer **les écarts à l'indépendance**.

Calcul des écarts à l'indépendance : $\delta_{ij} = (N_{ij} - N^*_{ij})$

N^*_{ij}	**Femme**	**Homme**	**Total**
Groupe 1	5.7	11.3	17
Groupe 2	6.3	12.7	19
Total	12	24	36

$N_{ij} - N^*_{ij}$	**Femme**	**Homme**	**Total**
Groupe 1	-2.7	+2.7	0
Groupe 2	+2.7	-2.7	0
Total	0	0	0

Entre variables discrètes

Test de significativité

Le test le plus commun est le test du *Chi-2*, quantifiant la somme des déviations entre effectifs observés et effectifs théoriques.

Calcul des *Chi-2* locaux : $\xi_{ij} = \frac{(N_{ij} - N^*_{ij})^2}{N^*_{ij}}$

Ces quantités sont des écarts relatifs. Plus le *Chi-2* local d'une case est élevé, plus la déviation entre valeurs observées et estimées est significative sur le plan statistique, et plus elle correspond à un évènement rare ayant peu de chance de se produire si X et Y était indépendant.

Calcul du *Chi-2* global : $\xi = \sum_{i=1}^{k} \sum_{j=1}^{\ell} \xi_{ij}$

où k et ℓ correspondent aux nombres de lignes et de colonnes du tableau de contingence. Plus ξ est grand, moins l'hypothèse d'indépendance est valide.

Entre variables discrètes

Test de significativité

Le nombre de **degré de liberté** est le nombre de cases pouvant produire des déviations indépendantes les unes des autres : $z = (k-1)(\ell-1)$

Reprenons notre exemple précédent :

ξ_{ij}	**Femme**	**Homme**	**Total**
Groupe 1	1.255	0.628	-
Groupe 2	1.129	0.561	-
Total	-	-	3.567

La déviation la plus significative concerne la sous-représentation des femmes dans le groupe 1. La valeur du *Chi-2* total du tableau vaut 3.567. Le nombre de degrés de liberté de ce tableau est (2-1)(2-1) soit 1 degré de liberté. Reste à faire un test d'indépendance selon un risque d'erreur α !

Conditions de validité du test du *Chi-2* :

- $N_{..} \geq 20$ $\forall i, j N_{i.} \geq 5, N_{.j} \geq 5$ $N^*_{ij} \geq 5$ dans 80% des cases du tableau de contingence