Analyse de données multidimensionnelles et Machine learning

Master 2ème année : Analyse des Systèmes Complexes

Gilles Roussel

EIL - Côte d'Opale

Année 2023-2024

Introduction

Ce cours présente :

- L'analyse d'un jeu de données par les statistiques descriptives (monovariée et bivariée)
- La réduction de dimension pour une analyse multivariée

On utilisera le langage Python comme outils de traitement. Les bases de ce langage sont supposées connues.

- Langage Python
- Les librairies suivantes :
  - Pandas
  - Matplotlib
  - Numpy
  - Scipy
- Notebook Jupyter

Un certain nombre d'éléments de vocabulaire sont à rappeler :

- **Individu** : des objets, des personnes, des animaux, des mesures physiques, etc.
- Des individus ont des caractéristiques : on les appelle des caractères, ou des **variables**.
- L'ensemble des individus s'appelle **la population**. On note souvent sa taille **N**. Il est très fréquent de ne pas connaître la taille exacte d'une population.
- Lorsque l'on sélectionne certains individus d'une population, on obtient **un échantillon**. Sa taille est souvent notée **n**.

Comment représenter les données ?

- On représente en général un échantillon sous forme de tableau, où chaque ligne correspond à un individu, et chaque colonne représente une variable.
- Cette représentation est à l'origine du format de fichier CSV (comma separated values). Ce format peut être ouvert avec les logiciels tableurs (Microsoft® Excel, OpenOffice Calc), et est facilement interprétable par les langages R et Python.
- Cette représentation est très similaire à celle des bases de données relationnelles.

identifiant_transaction	date_operation	date_valeur	libelle	debit	credit	solde
242	2023-10-06	2023-10-06	PRELEVEMENT XX TELEPHONE XX XX	-13,58		2513,79
69	2023-10-06	2023-10-06	CARTE XX XX CHEZ LUC XX	-10		2527,37
25	2023-10-06	2023-10-06	FORFAIT COMPTE SUPERBANK XX XX XX XX	-1,92		2537,37
299	2023-10-05	2023-10-05	CARTE XX XX XX XX XX XX	-10,64		2539,29
45	2023-10-05	2023-10-05	CARTE XX XX XX XX	-4,8		2549,93
95	2023-10-03	2023-10-03	CARTE XX XX LA CCNCF XX	-67,68		2554,73
78	2023-10-02	2023-10-02	VIREMENT XX XX XX XX XX XX XX XX XX XX XX XX		676	2622,41

Figure 1: Représentation d'un échantillon

Statistique et probabilités: quelle différence ?

- Quand on ne fait qu'observer et décrire objectivement un phénomène, alors on fait des **statistiques**.
- Mais dès lors que l'on modélise, on fait le lien entre ce qu'on observe et le domaine théorique que constituent les **probabilités** : on passe alors dans le domaine de la **statistique dite inférentielle**.
- Exemple :
  - Si vous étudiez la proportion femmes/hommes d'un pays, vous sélectionnez un échantillon dans lequel vous observez ces proportions : par exemple 55% de femmes et 45% d'hommes. Ce sont des statistiques.
  - Mais si vous dites ensuite dans ce pays, un enfant qui naît a une probabilité de 55% d'être une fille, alors vous faites des probabilités !

Les différents domaine de la statistique sont :

- **Les statistiques descriptives :** présenter, décrire et résumer le jeu de données, à l'aide de graphiques et de mesures (moyenne, écart-type, etc.). En statistique descriptive, chaque graphique (ou chaque mesure) est calculé(e) sur 1 ou 2 variables à la fois, pas plus.

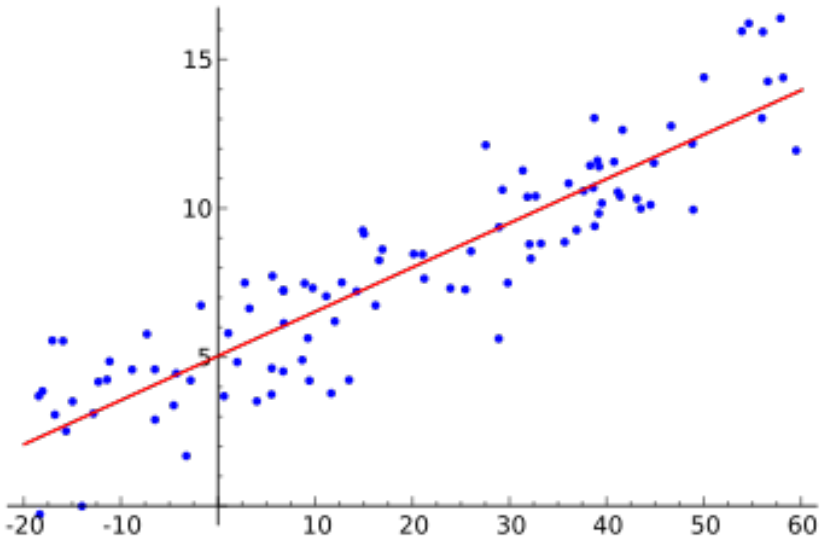


Figure 2: Graphique en 2 dimensions avec 1 axe horizontal et 1 axe vertical (source: Wikipedia)

- **L'analyse exploratoire des données :** prolongement des statistiques descriptives, en s'intéressant aux relations entre 3 variables ou plus. Représenter des graphiques avec 3, 4, 5 ou 100 dimensions n'est plus possible sur du papier à 2 dimensions.
- **Les statistiques inférentielles** : il s'agit d'analyser les données d'un sous-ensemble d'une population pour en déduire les caractéristiques globales de la population. On met en oeuvre alors des estimateurs ou de tests statistiques.
- **La modélisation statistique :** Il s'agit d'observer les caractéristiques d'un échantillon, puis de formaliser ces observations par des règles mathématiques. Cette formalisation s'appelle un **modèle probabiliste**. Une fois que l'on a décrit un phénomène par un modèle, on peut faire de la **prédiction** ou de la **prévision**.

Tout çela est du **"Data Analysis"** au sens anglophone du terme. Un **Data Scientist** doit maîtriser les 3 premiers domaines mais doit également savoir **modéliser** un phénomène.

Partie I

Statistiques descriptives

Les données de travail

Fichiers utiles

Dans cette partie, on va utiliser un jeu de données permettant d'analyser les opérations d'un relevé de banque. Déposer et décompresser l'archive *Analyse* dans votre espace de travail. Il contient des fichiers aux formats :

- **.csv** : les données que nous utiliserons pour illustrer ce cours.
- **.py** : code informatique en langage Python.
- **.ipynb** : code informatique en langage Python, dans le format du notebook Jupyter.
- On se propose d'analyser les opérations contenues dans le fichier "operations.csv".
- Une ligne = une opération = un individu de la population des opérations.

Le fichier CSV ouvert avec un éditeur de texte ressemble à ceci :

- identifiant_transaction,date_operation,date_valeur,libelle,debit,credit,solde
- 242,2023-10-06,2023-10-06,FORFAIT COMPTE SUPERBANK XX XX XX XX,-1.92„2513.79
- 69,2023-10-06,2023-10-06,CARTE XX XX CHEZ LUC XX,-10.00„2515.71

Importons dans un premier temps l'ensemble des librairies qui vont nous servir durant l'entièreté de ce cours:

- *import pandas as pd*
- *import numpy as np*
- *import matplotlib.pyplot as plt*
- *import seaborn as sns*
- *import datetime as dt*
- *import scipy.stats as st*

Nous pouvons à présent charger le jeu de données dans un dataframe (df) que nous nommerons ici data. Nous affichons ensuite les 5 premières lignes.

- *data = pd.read_csv('operations.csv')*
- *data.head()*

	date_operation	libelle	montant	solde_avt_ope	categ
0	2023-03-31	DON XX XX XX XX XX XX XX	-1.44	1515.25	AUTRE
1	2023-04-03	CARTE XX XX RAPT XX	-24.00	1513.81	TRANSPORT
2	2023-04-03	CARTE XX XX RAPT XX	-73.00	1489.81	TRANSPORT
3	2023-04-03	VIREMENT XX XX XX XX XX XX XX XX XX XX XX	676.00	1416.81	AUTRE
4	2023-04-03	VIREMENT XX XX XX XX XX XX	4.80	2092.81	AUTRE

Figure 3: Extraction de 5 lignes du tableau 'operations'

Les données de travail

Les variables 'operations.csv'

La commande *data.shape* permet de connaître les dimensions du tableau
out : (309, 5)
Nous avons donc 309 transactions, renseignées sur 5 variables différentes :

- date_operation : date de l'opération
- libelle : libellé de l'opération
- montant : montant de l'opération
- solde_avt_ope : solde du compte avant l'opération considérée
- categ : catégorie d'achat

Nettoyage du jeu de données

Erreur de type

Plusieurs erreurs se sont glissées dans ce jeu de données. Commencçons par résoudre les erreurs de type.
Les variables ont-elles été importées dans le bon type ?

- *data.dtypes*
- Out[] :
  - date_operation: object
  - libelle: object
  - montant: float64t
  - solde_avt_ope: float64
  - categ: object
  - dtype: object

La variable date n'est pas considérée comme une date. On peut corriger cela facilement via la fonction to_datetime de pandas.

- *data['date_operation'] = pd.to_datetime(data['date_operation'])*

Nettoyage du jeu de données

Valeurs manquantes

On souhaite vérifier si notre jeu de données contient des valeurs manquantes :

- *data.isnull().sum()*
- Out[] :
  - date_operation 0
  - libelle 0
  - montant 2
  - solde_avt_ope 0
  - categ 1
  - dtype: int64

pour afficher uniquement les variables qui ont des valeurs manquantes :

- *nb_na = data.isnull().sum()*
- *nb_na[nb_na>0]*
- Out [] :
  - montant 2
  - categ 1
  - dtype: int64

On observe que notre jeu de données contient 3 valeurs manquantes.

Regardons les plus en détails les valeurs manquantes :

- *data.loc[data['montant'].isnull(),:]*
- Out[] :

	date_operation	libelle	montant	solde_avt_ope	categ
107	2023-06-12	CARTE XX XX LES ANCIENS ROBINSON XX	NaN	4667.19	COURSES
269	2023-09-11	CARTE XX XX XX XX	NaN	3401.93	AUTRE

Ici, il reste relativement simple de remplacer les valeurs manquantes. Le montant manquant correspond donc au solde de l'opération suivante, moins le solde de l'opération concernée :

- *data_na = data.loc[data['montant'].isnull(),:]* # on stocke le df des valeurs manquantes dans un nouveau df
- *for index in data_na.index:* # pour chaque ligne de mon df, on récupère les index (qui ne changent pas au travers du .loc)
- *data.loc[index, 'montant'] = data.loc[index+1, 'solde_avt_ope'] - data.loc[index, 'solde_avt_ope']* # calcul du montant à partir des soldes précédents et actuels

Nettoyage du jeu de données

Valeurs manquantes (suite)

A présent regardons la catégorie manquante :

- *data.loc[data['categ'].isnull(),:]*

	date_operation	libelle	montant	solde_avt_ope	categ
156	2023-07-06	PRELEVEMENT XX TELEPHONE XX XX	-36.48	3295.68	NaN

Par manque d'informations, on devrait supprimer ici la ligne correspondante. Mais regardons si nous ne pouvons pas trouver la catégorie à partir des autres informations, notamment le libellé :

- data.loc[data['libelle'] == 'PRELEVEMENT XX TELEPHONE XX XX', :]

	date_operation	libelle	montant	solde_avt_ope	categ
8	2023-04-05	PRELEVEMENT XX TELEPHONE XX XX	-7.02	2056.02	FACTURE TELEPHONE
62	2023-05-09	PRELEVEMENT XX TELEPHONE XX XX	-7.02	4090.10	FACTURE TELEPHONE
102	2023-06-07	PRELEVEMENT XX TELEPHONE XX XX	-6.38	4688.91	FACTURE TELEPHONE
156	2023-07-06	PRELEVEMENT XX TELEPHONE XX XX	-36.48	3295.68	NaN
204	2023-08-07	PRELEVEMENT XX TELEPHONE XX XX	-7.46	3751.73	FACTURE TELEPHONE
260	2023-09-05	PRELEVEMENT XX TELEPHONE XX XX	-6.38	3453.96	FACTURE TELEPHONE
308	2023-10-06	PRELEVEMENT XX TELEPHONE XX XX	-13.58	2413.58	FACTURE TELEPHONE

On déduit assez facilement que la catégorie manquante ici est : FACTURE TELEPHONE

- data.loc[data['categ'].isnull(), 'categ'] = 'FACTURE TELEPHONE'

Nettoyage du jeu de données

Doublons

Regardons à présent si certaines transactions sont apparues en doublons. Pour cela, on se concentrera sur des informations qui ne peuvent normalement pas être doublés, soit : la date, le libelle, le montant et le solde avant opération. Sur ces 4 variables, il n'est normalement pas possible d'avoir deux transactions identiques :

- data.loc[data[['date_operation', 'libelle', 'montant', 'solde_avt_ope']].duplicated(keep=False),:]

	date_operation	libelle	montant	solde_avt_ope	categ
43	2023-04-25	CARTE XX XX LES ANCIENS ROBINSON XX	-32.67	3647.67	COURSES
44	2023-04-25	CARTE XX XX LES ANCIENS ROBINSON XX	-32.67	3647.67	COURSES

On a ici une opération qui est complètement en double. Il suffit donc de supprimer l'une des deux via le drop_duplicate

- data.drop_duplicates(subset=['date_operation', 'libelle', 'montant', 'solde_avt_ope'], inplace=True, ignore_index=True)

Un describe peut potentiellement nous aider dans un premier temps :

- data.describe()

	montant	solde_avt_ope
count	308.000000	308.000000
mean	-45.782013	3395.301071
std	872.818105	667.109412
min	-15000.000000	1416.810000
25%	-20.447500	3010.737500
50%	-9.600000	3452.465000
75%	-2.715000	3787.232500
max	1071.600000	4709.310000

Nettoyage du jeu de données

Détection d'outliers

Une bonne première approche en attendant d'avoir des outils plus adéquats, est de regarder le maximum et le minimum. Cela donne généralement un premier apperçu de ce qui pourrait clocher à ce niveau. Ici on voit un minimum de montant de -15 000 (ce qui correspondrait à un débit de -15000). Cela semble assez étonnant, d'autant que le solde ne semble pas descendre en conséquence à aucun moment (le max est 4700 et le minimum 1416). Vérifions les soldes autour de cette transaction :

- i = data.loc[data['montant']==-15000,:].index[0] # récupération de l'index de la transaction à -15000
- data.iloc[i-1:i+2,:] # on regarde la transaction précédente et la suivante

	date_operation	libelle	montant	solde_avt_ope	categ
197	2023-08-03	VIREMENT XX XX XX XX XX XX XX XX XX XX XX	676.00	3121.35	AUTRE
198	2023-08-03	CARTE XX XX XX XX	-15000.00	3797.35	AUTRE
199	2023-08-03	CARTE XX XX L'EPICERIE DEMBAS XX XX	-10.51	3782.96	AUTRE

Il y a en effet une grosse incohérence. Les soldes nous indique une opération de -14.39 et non de -15000. Il y a en effet une valeur abérrante ici ! Remplaçons donc là pour sa valeur initiale :

- data.loc[data['montant']==-15000, 'montant'] = -14.39

Le jeu de données semble à présent propre, on peut avancer sur la partie analyse.

Les données de travail

Fichiers utiles

On peut procéder à l'enrichissement du fichier précédent par d'autres variables (déduites ou pas). Nouveau fichier : *"operations_enrichies.csv"* (à mettre dans le répertoire de travail). Il contient les variables :

- *sens* : indique si l'opération est un crédit ou un débit
- *solde_avt_operation* : elle indique le solde restant avant que l'opération ne soit effectuée
- *categ* : qui indique la catégorie de l'opération : "courses", "loyer", "facture", etc. (à définir "à la main")
- *type* : indiquant le type d'opération : "virement", "paiement par carte", 'retrait", etc.
- *tranche_depense* : si l'opération est une petite, moyenne, etc.
- *annee* : l'année, déduite de date_operation
- *mois* : le mois, déduit de date_operation
- *jour* : le jour du mois (compris entre 1 et 31)
- *jour_sem* : le jour de la semaine (lundi, mardi, etc.)
- *jour_sem_num* : numéro du jour semaine (compris entre 1 et 7)
- *weekend* : indique si la date d'opération se situe sur un weekend
- *quart_mois* : indique l'avancée dans le mois (1 : début, ..., 4 : fin de mois)
- *attente* : durée (en jours) séparant 2 opérations de catégorie "courses". Un truc à estimer par un algo ...

Les données de travail

Fichiers utiles : vous pouvez utiliser vos propres relevés de compte!

En général, si vous avez accès à vos comptes bancaires à partir d'internet, il existe un endroit où vous pouvez télécharger le relevé de vos opérations. Si vous avez le choix, choisissez le format CSV, et téléchargez le plus d'opérations possibles. Si seuls les formats de fichiers XLS ou ODS vous sont proposés, alors il faudra ouvrir le fichier dans un tableur (Excel ou OpenOffice Calc), puis l'enregistrer sous un format CSV. Pour plus de précision

Types de variables

Il y a 2 types de variables sub-divisés en 2 groupes.

- Les variables **quantitatives** (ex: debit, credit, date, solde, année) pouvant être :
  - **continues** (ou considérées comme telles) si la résolution est très fine. Ex : 0.01, 0.02, 0.03 ... 10000.00)
  - **discrètes** définies sur un ensemble
- Les variables **qualitatives** nominales ou ordinales.
  - Une variable est **ordinale** si ses modalités peuvent être ordonnées. Ex: "tranche_depense" est **ordinale** :
    «petite dépense» $<=$ «dépense moyenne» $<=$ «grosse dépense».
  - Une variable est **nominale** sinon

Représenter la distrib. empir. d'une variable (1/2)

Les données de la variable *categ* d'un tableau de 1000 individus n'est pas facile à visualiser et à interpréter.

- Les différentes "possibilités" que l'on peut observer pour la variable *categ* sont ses **modalités**. Ex : courses, transport, autre, loyer, etc.
- On associe à chaque modalité (ou valeur) un **effectif**. L'effectif de la modalité courses est ncourses=39.
- En divisant un effectif par le nombre d'individus de l'échantillon (noté n), on obtient une **fréquence**.
- La **distribution empirique** d'une variable est l'ensemble des valeurs (ou modalités) prises par cette variable, ainsi que son effectif associé et/ou sa fréquence.

modalité	effectif	fréquence
AUTRE	212	0.688312
COURSES	39	0.126623
TRANSPORT	21	0.068182
[...]	[...]	[...]

Figure 4: Distribution empirique

Représenter la distrib. empir. d'une variable (1/2)

Code

VARIABLE QUALITATIVE
On commence par sélectionner la colonne souhaitée :
data['categ'],
puis on compte le nombre d'apparitions de chaque modalité :
data['categ'].value_counts()
Pour obtenir les fréquences, on peut éventuellement ajouter : *normalize=True*.
On obtient donc la distribution empirique. Pour l'afficher, on fait appel à la méthode *plot*,
à laquelle on spécifie le type de graphique souhaité (pie ou bar).
Diagramme en secteurs
data["categ"].value_counts(normalize=True).plot(kind='pie')
Cette ligne assure que le pie chart est un cercle plutôt qu'une ellipse
plt.axis('equal') plt.show() # Affiche le graphique
Diagramme en tuyaux d'orgues data["categ"].value_counts(normalize=True).plot(kind='bar')
plt.show()

Représenter la distrib. empir. d'une variable (2/2)

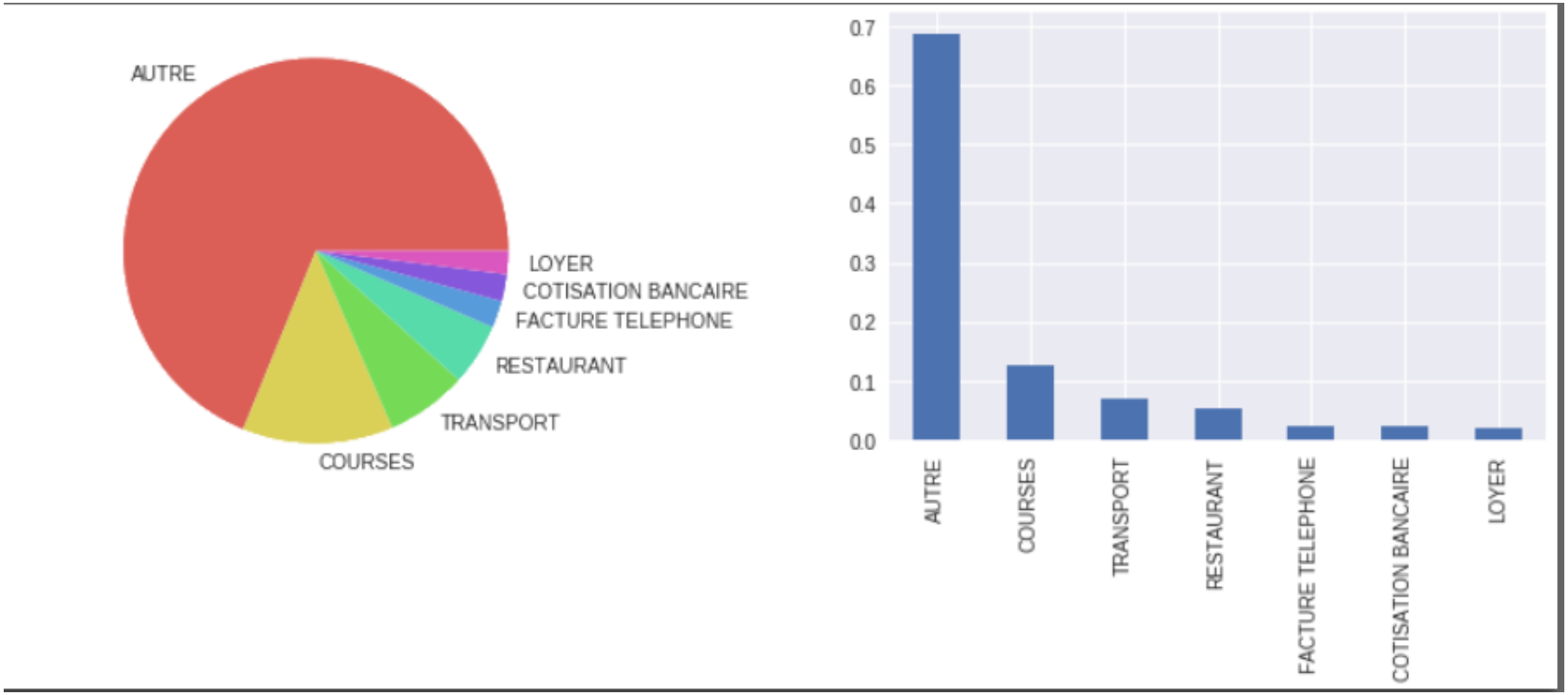


Figure 5: Distribution empirique : diagramme camembert, diagramme bâtons d'orgue, (pie chart, bar chart)

Agrégation de classe

Variables continues

- Exemple de la taille d'une personne, on peut avoir une personne de taille 1,47801 m et une autre de 1,47802 m. Considérer que 1,47801 m et 1,47802 m sont presque égales, c'est regrouper ces valeurs. On dit alors que l'on agrège des valeurs en classes.
- Si on décide d'agréger en classes de taille 0,2 m, alors ces 2 valeurs seront toutes les deux situées dans la classe [1.4m;1.6m[
- Le fait d'agréger une variable s'appelle la discrétisation (en anglais : binning, bucketing ou discretization).

- On utilise l'histogramme, dans lequel les valeurs sont agrégées. La largeur des rectangles correspondent à la largeur de la classe.
- Il est possible d'agréger en classes de largeurs inégales. Ex: classes de largeur 0,5 m pour les tailles inférieures à 1 m, classes de largeur 0,2 m pour les tailles supérieures à 1 m. On a alors : [0m;0.5m[, [0.5m;1m[, [1m;1.2m[, [1.2m;1.4m[

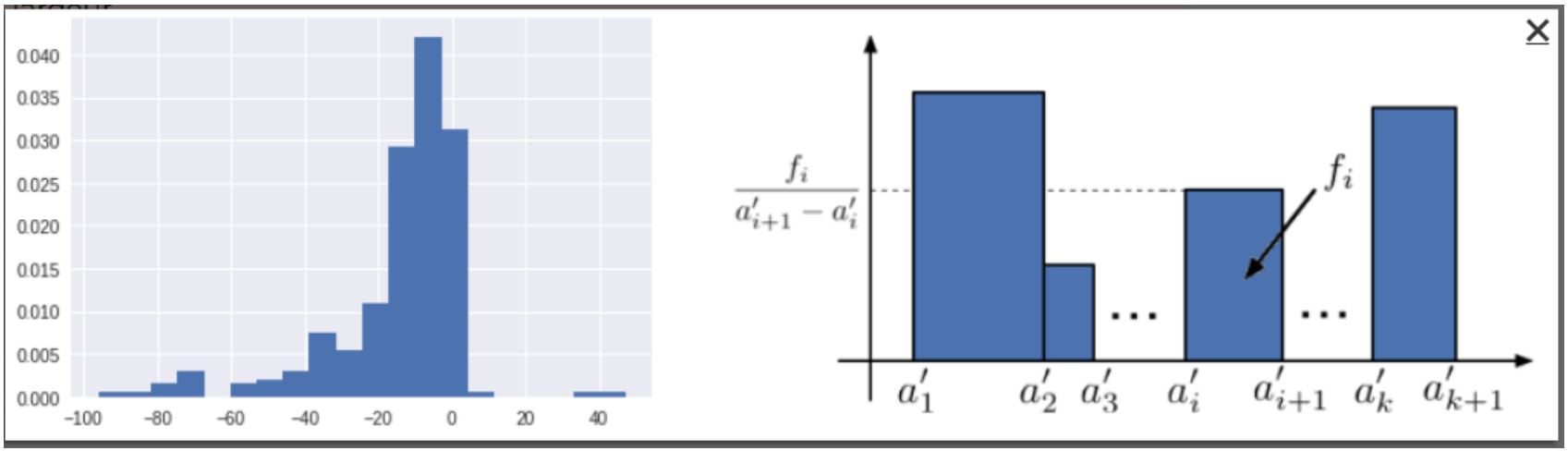


Figure 6: histogramme

- Si l'on ne souhaite pas devoir agréger les données, on peut faire appel à la fonction de répartition empirique.
- On parcourt l'axe horizontal des petites valeurs vers les grandes valeurs. A chaque fois que l'on rencontre une valeur présente dans l'échantillon, on monte d'une marche. Il y a donc autant de marches que de valeurs, donc d'individus.
- Si il y a 2 valeurs égales ou plus, alors la marche est d'autant plus grande, assez rare pour les valeurs continues.
- En normalisant, la hauteur totale de l'escalier vaut 1.
- $F_{emp}(x) = \frac{1}{n}\sum_{i=1}^{n} I_{x_i \leq x}$ avec la fonction indicatrice $I_{x_i \leq x} = 1$ si $x_i \leq x$, =0 sinon
- Le nombre optimal de classes vaut (Sturges, 66) : $k = [1 + log_2(n)]$

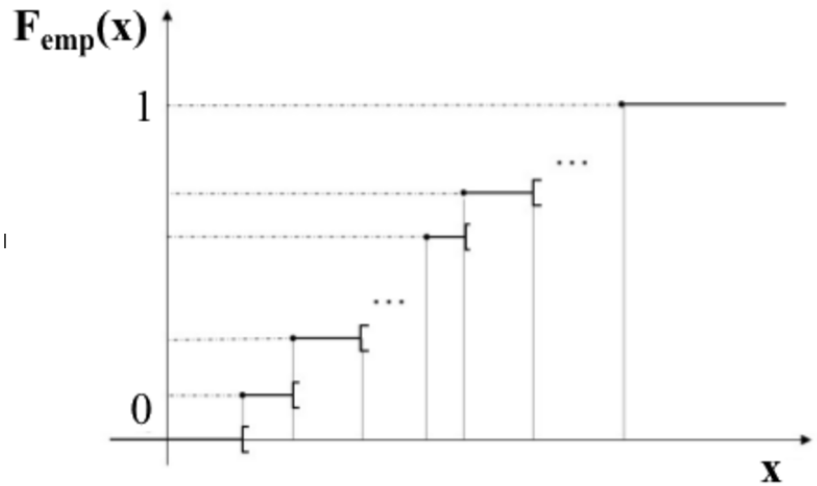


Figure 7: Fonction de répartition

Représenter les variables sous forme de tableau

Variables qualitatives

- Pour les variables qualitatives, il suffit juste de compter le nombre de valeurs pour chaque modalité. Ce nombre est appelé *effectif* de la modalité. Ainsi, pour une modalité *ai* (i est compris entre 1 et k), on note l'effectif *ni*.
- Si on divise l'effectif par n (la taille de l'échantillon), on obtient la fréquence compris entre 0 et 1. La somme des fréquences donne 1 !
- Avec l'exemple de la variable *categ* :

X	n	f
a_1	n_1	f_1
$\vdots$	$\vdots$	$\vdots$
a_k	n_k	f_k

categ	n	f
AUTRE	212	0.688312
COURSES	39	0.126623
TRANSPORT	21	0.068182
RESTAURANT	16	0.051948
FACTURE TELEPHONE	7	0.022727
COTISATION BANCAIRE	7	0.022727
LOYER	6	0.019481

Figure 8: Table des effectifs de variables qualitatives

Représenter les variables sous forme de tableau

Variables quantitatives et fréquences cumulées

- Pour les variables quantitatives discrètes, on peut ajouter une colonne qui donne la fréquence cumulée.
- La fréquence cumulée d'une modalité a_i, est la somme des fréquences de toutes les modalités inférieures ou égales à a_i ou de la classe $[a'_i, a'_{i+1}[$. On la note F

X	n	f	F
a_1	n_1	f_1	$F_1 = f_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_i	n_i	f_i	$F_i = f_1 + \ldots + f_i$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_k	n_k	f_k	$F_k = 1$

quart_mois	n	f	F
1	86	0.279221	0.279221
2	76	0.246753	0.525974
3	75	0.243506	0.769481
4	71	0.230519	1.000000

Figure 9: Table des effectifs cumulés de variables quantitatives discrètes

X	n	f	F
$[a'_1, a'_2[$	n_1	f_1	$F_1 = f_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[a'_i, a'_{i+1}[$	n_i	f_i	$F_i = f_1 + \ldots + f_i$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[a'_k, a'_{k+1}[$	n_k	f_k	$F_k = 1$

montant	n	f	F
[...]	[...]	[...]	[...]
[-120.0 ; -90.0 [	2	0.006494	0.048701
[-90.0 ; -60.0 [	11	0.035714	0.084416
[-60.0 ; -30.0 [	28	0.090909	0.175325
[-30.0 ; 0.0 [	237	0.769481	0.944805
[0.0 ; 30.0 [	3	0.009740	0.954545
[...]	[...]	[...]	[...]

Figure 10: Table des effectifs cumulés de variables quantitatives continues

Analyse monovariée

- **Une statistique** = indicateur numérique, plus ou moins efficace, calculé à partir d'un échantillon.
- Elle nous permet de résumer un grand échantillon en un seul nombre.
- **Remarque** : On peut calculer le taux de réussite à partir des réponses des étudiants, mais on ne peut pas retrouver les réponses des étudiants uniquement avec le taux de réussite !
- A la différence d'un indice, un indicateur est très neutre, comme une moyenne par exemple.
- Dans ce chapitre, nous allons effectuer des analyses univariées. Une analyse univariée est une analyse effectuée sur une variable à la fois.

- Exemple. Vous demandez à un ami : Combien de temps dure le trajet entre les deux villes ?
- Il vous répond : "La plupart du temps, je mets entre 40 min et 45 min". il vous donne ici une **mesure de tendance centrale** qui s'appelle **le mode**.
- **Pour les variables qualitatives**, ou pour les variables quantitatives discrètes, le mode est la modalité ou la valeur la plus fréquente.
- **Pour les variables quantitatives** continues, on travaille dans le cas agrégé, en regroupant les valeurs par classes. **La classe modale** est la classe la plus fréquente.
- Dans sa réponse précédente, votre ami a découpé sa variable en tranches de 5 minutes et a déterminé que la tranche la plus fréquente était [40min;45min[

- Vous demandez encore à votre ami : Oui, mais je ne peux pas me contenter de la durée la plus fréquente : car si la deuxième durée la plus fréquente est de 65 à 70 minutes, il faut que je parte beaucoup plus tôt !
- Il répond alors : Oui tu as raison. En fait je mets en moyenne 60 minutes par trajet, car il y a souvent des embouteillages.
- La moyenne coïncide avec le centre de gravité.
- Parfois les valeurs les plus fréquentes ne sont pas très près de la moyenne, à cause des outliers qui sont très influant sur la moyenne.

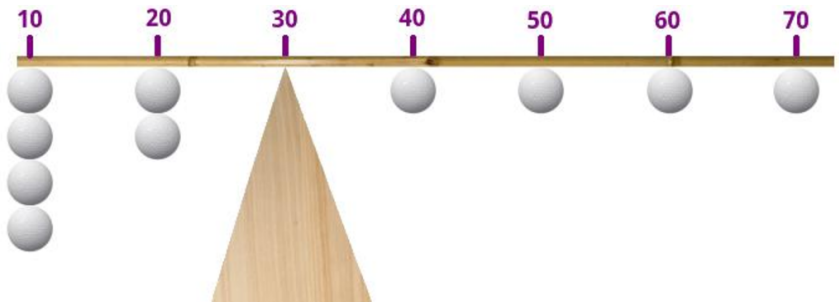
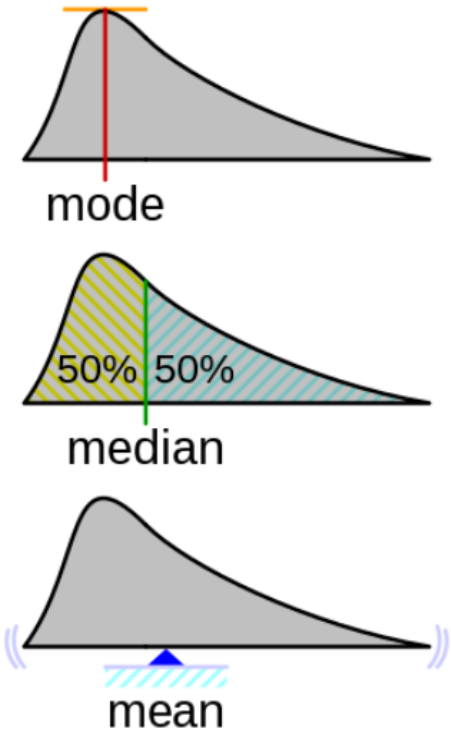


Figure 11: Moyenne et Centre de gravité : les outliers sont impactant

- Vous demandez à votre ami : Quand tu me dis que tu mets en moyenne 60 min, j'imagine que tu considères dans ton calcul les rares fois durant lesquelles il y avait de la neige, et que tu as mis 4 h à faire la route ?
- Il répond : Oui effectivement. Disons que la moitié des trajets que j'ai effectué ont pris plus de 55 min, et l'autre moitié ont pris moins de 55 min.
- La médiane, (notée Med), est la valeur telle que le nombre d'observations supérieures à cette valeur est égal au nombre d'observations inférieures à cette valeur.
- Pour trouver la médiane de n valeurs, il faut commencer par les trier. Une fois triées, on appelle x(1) la première valeur, x(2) la deuxième valeur, ... , et x(n) la dernière valeur. La médiane, c'est la valeur qui sera exactement au milieu du classement, soit :
$Med = x((n+1)/2)$

Pour un histogramme,

- Le mode est le "point le plus haut" de la distribution,
- La médiane est la valeur qui divise la surface en deux,
- La moyenne est le centre de gravité de la distribution.



(Source : commons.wikimedia.org, licence GFDL)

Figure 12: Tendances centrales et histogramme

Dans l'exemple précédent du temps de trajet, votre ami vous a donné des mesures de tendance centrale, comme par exemple la moyenne, qui est de 60 minutes par trajet.

- Ce qui manque maintenant, c'est de savoir si les durées que votre ami a effectué sont très "resserrées" autour de 60 min (exemple : [58, 60, 62, 59, 57, ...]), ou bien si elles s'en écartent beaucoup (exemple : [40, 70, 78, 43, ...]).
- Si les valeurs sont très resserrées autour de 60 minutes, alors prévoyez de partir 75 minutes à l'avance. Ainsi, il est probable que vous arriverez 5 ou 10 minutes avant votre entretien. Mais si les valeurs sont très écartées, alors prévoyez plutôt de partir 100 minutes à l'avance, car il est tout à fait possible que le trajet dure 80 minutes !
- La variance empirique est l'indicateur le plus utilisés en statistiques pour évaluer la dispersion
- $v = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Une forme pratique (König-Huygens) : $v = (\frac{1}{n}\sum_{i=1}^{n} x_i^2) - \bar{x}^2$

- L'écart-type empirique, est la racine carrée de la variance empirique.
- $s = \sqrt{v}$

Variance corrigée

La meilleure manière d'estimer la variance d'une variable aléatoire (i.e. la variance théorique) n'est pas d'utiliser la variance empirique.

- La variance empirique donne des valeurs qui (en moyenne) sont inférieures à la variance de la variable aléatoire.
- La variance empirique est un estimateur biaisé de la variance de la variable aléatoire.
- La variance empirique sans biais est égale : $s'^2 = \frac{n}{n-1} v$, où v est la variance empirique.

- Un écart-type de 6,55 minutes sur un trajet de 1 h, ce n'est pas la même chose qu'un écart-type de 6,55 minutes sur un trajet de 24 h !
- Pour rendre compte de cela, on a créé le **coefficient de variation**, qui est l'écart-type empirique divisé par la moyenne :

$$CV = \frac{s}{\bar{x}}$$

- Une autre mesure de dispersion, **l'écart moyen absolu** :

$$EMA = \frac{1}{n}\sum_{i=1}^{n} |x_i - Med|$$

Analyse monovariée : Les mesures de dispersion

L'écart inter-quartiles

- Un quartile, est la valeur en dessous de laquelle on trouve un quart des valeurs de l'échantillon.
- Il existe 3 quartiles, notés Q1 (premier quartile), Q2 (deuxième quartile=médiane) et Q3 (troisième quartile).
- 1/4 des valeurs se trouvent en dessous de Q1 et 3/4 au dessus
- Il y a également les déciles (quantiles d'ordre 0.1, 0.2, etc.), ou les centiles, aussi appelés percentiles (quantiles d'ordre 0,01, 0.02, etc.)

- Elle permet de représenter schématiquement une distribution, en incluant sa dispersion.
- La boîte est délimitée par Q1 et Q3 , et on représente souvent la médiane à l'intérieur de la boîte.
- On dessine ensuite des moustaches à cette boîte, qui vont de la valeur minimale (A) à la valeur maximale (B)... à condition que la moustache (d'un côté ou de l'autre) ne mesure pas plus de 1,5 fois l'écart inter-quartiles *IQ*

$$IQ = Q3 - Q1$$

.

- Si certaines valeurs sont au dessous de $Q1 - 1.5xIQ$ ou au dessus de $Q3 + 1.5xIQ$, alors on les considère comme des outliers, et on ne les inclut pas dans la moustache

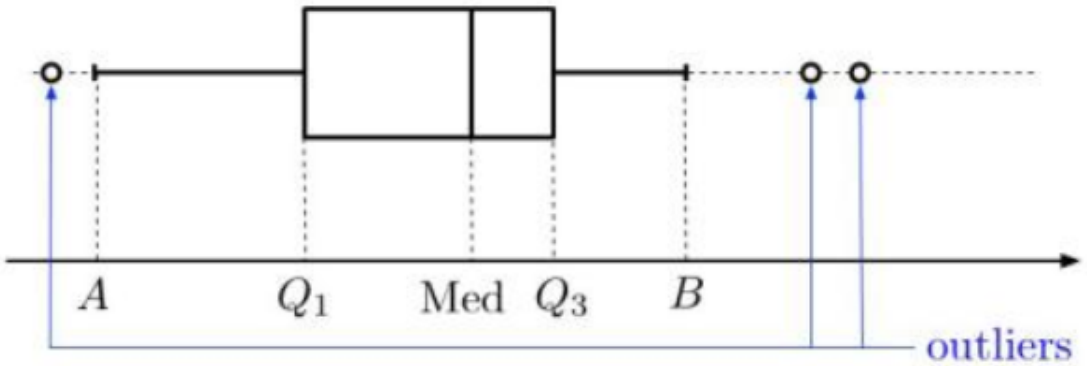


Figure 13: Boite à moustaches

- Votre ami vous a donné la moyenne des temps de trajets, ainsi que l'écart-type. Vous êtes déjà plus serein. Mais... il y a quelque chose que vous n'avez pas prévu. Regardez ces 2 distributions :

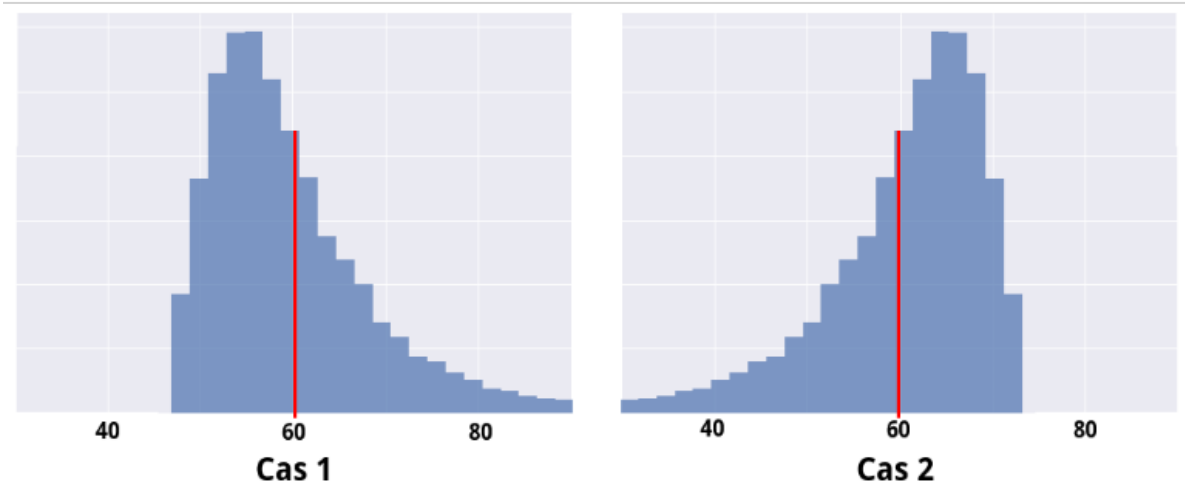


Figure 14: Formes d'histogramme

- Elles ont la même moyenne empirique (60 minutes), et le même écart-type. Cependant, le cas 1 est plus "risqué" que le cas 2.
- Dans le cas 2, il est très peu probable que votre trajet dure plus de 75 minutes : pas de risque d'être en retard !
- Par contre, dans le cas 1, il est tout à fait possible que votre trajet dure 80 minutes, ou même beaucoup plus.

- Le skewness est un moment qui mesure l'asymétrie.
- $\gamma_1 = \frac{\mu_3}{s^3}$ avec $\mu_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$
- L'asymétrie d'une distribution traduit la régularité (ou non) avec laquelle les observations se répartissent autour de la valeur centrale. On interprète cette mesure de cette manière :
- Si $\gamma_1 = 0$ alors la distribution est symétrique.
- Si $\gamma_1 > 0$ alors la distribution est étalée à droite.
- Si $\gamma_1 < 0$ alors la distribution est étalée à gauche.

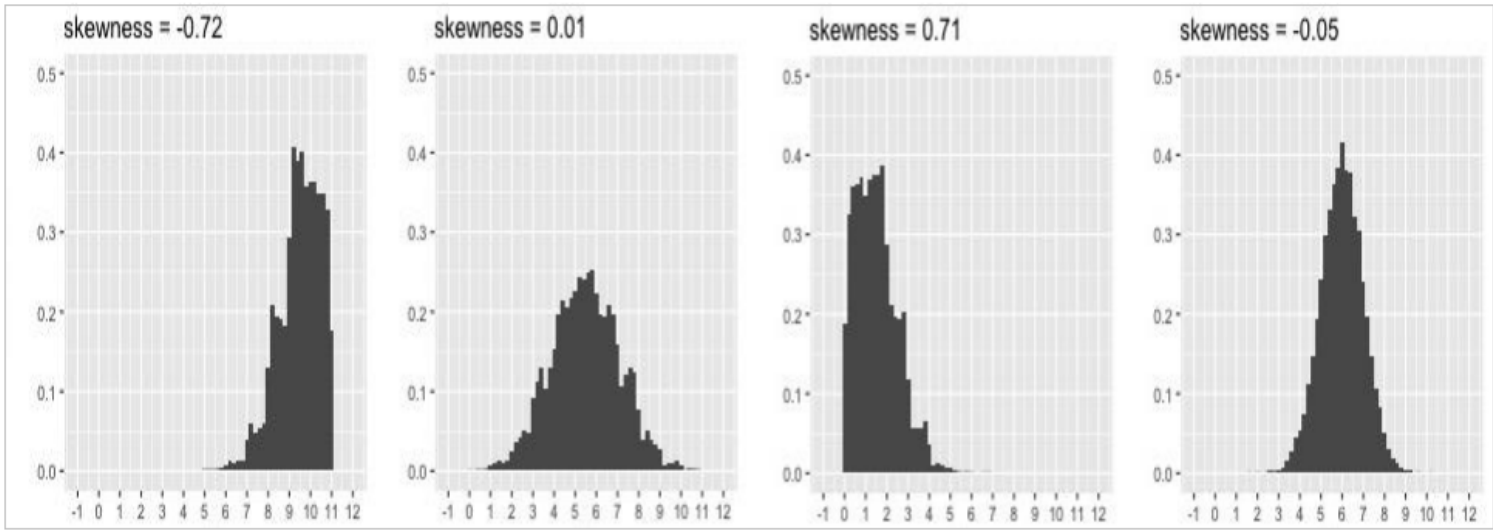


Figure 15: Skewness

- Comme le skewness, le kurtosis γ_2 est un moment
- γ_2 mesure l'aplatissement.
- $\gamma_2 = \frac{\mu_4}{s^4}$ avec $\mu_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$
- L'aplatissement peut s'interpréter à la condition que la distribution soit symétrique.
- On compare l'aplatissement par rapport à la distribution la distribution normale ("courbe de Gauss" ou "Gaussienne")

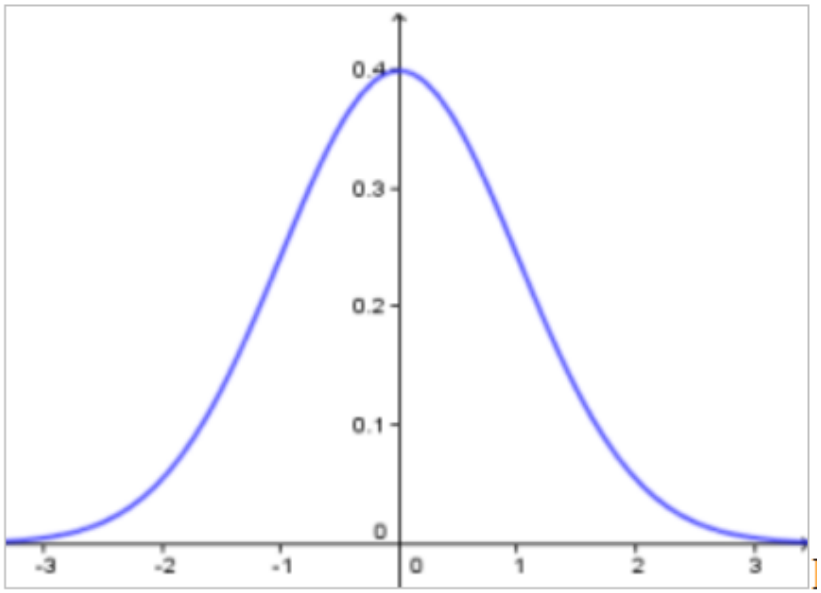


Figure 16: Loi normale

- Le kurtosis γ_2 s'interprète comme ceci :
- Si $\gamma_2 = 0$ alors la distribution a le même aplatissement que la loi normale
- Si $\gamma_2 > 0$ alors la distribution est moins aplatie que la loi normale. les observations sont plus concentrées.
- Si $\gamma_2 < 0$ alors la distribution est plus aplatie que la loi normale. les observations sont moins concentrées.

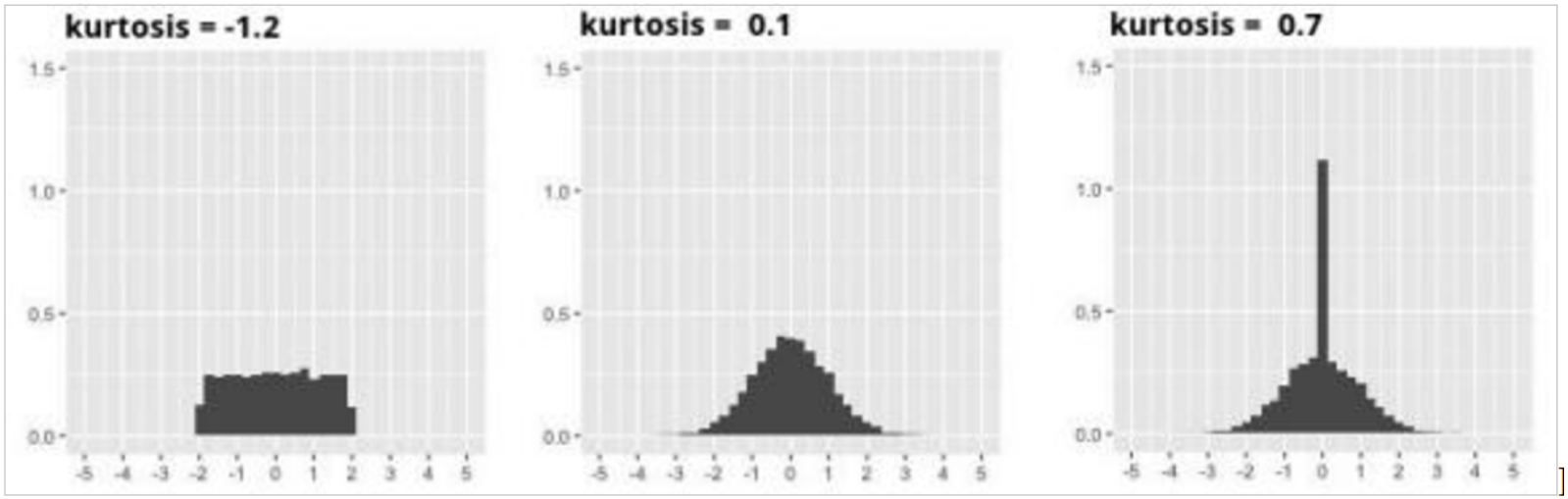


Figure 17: Loi normale

Les mesures de concentration sont le plus souvent utilisées pour des sommes d'argent ! Étudier la concentration d'argent, c'est regarder si l'argent est répartie de manière égalitaire ou pas. Pour visualiser cela, nous utilisons la courbe de Lorenz.

- Il faut vous imaginer la courbe de Lorenz comme un podium, non pas avec 3 places, mais avec autant de places que de gens. Ce podium ressemble à un escalier, sur lequel on place l'individu qui gagne le plus d'argent tout en haut, et celui qui gagne le moins d'argent tout en bas.

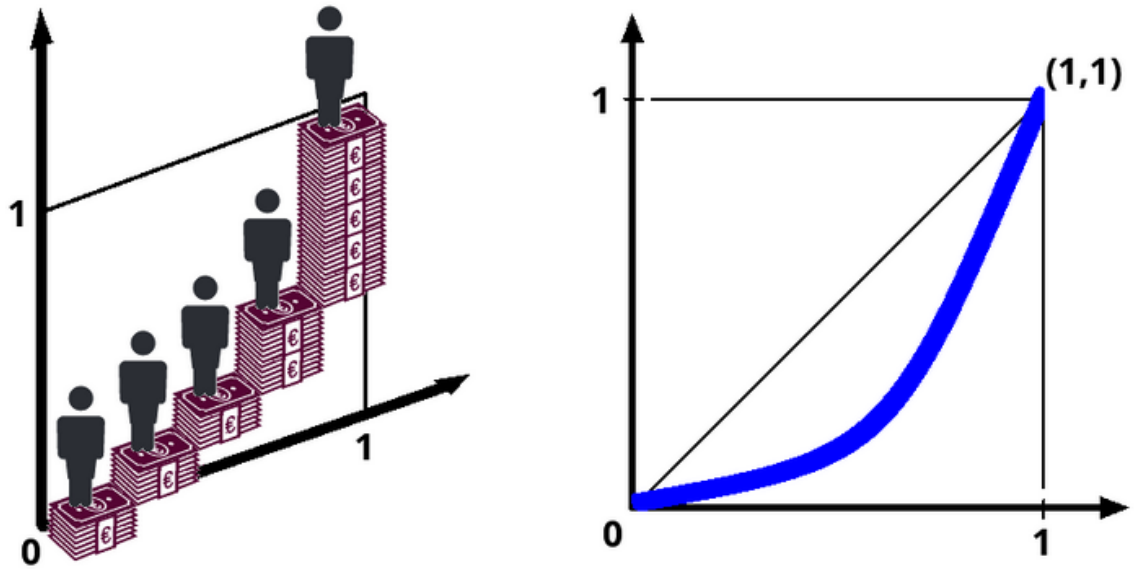


Figure 18: Courbe de Lorenz : répartition concentrée

Dans ce cas, la répartition est la plus égalitaire possible. L'escalier se présente comme ceci (à gauche) :

- On voit que les marches sont régulières, et que toutes les personnes sont alignées sur une droite appelée première bissectrice

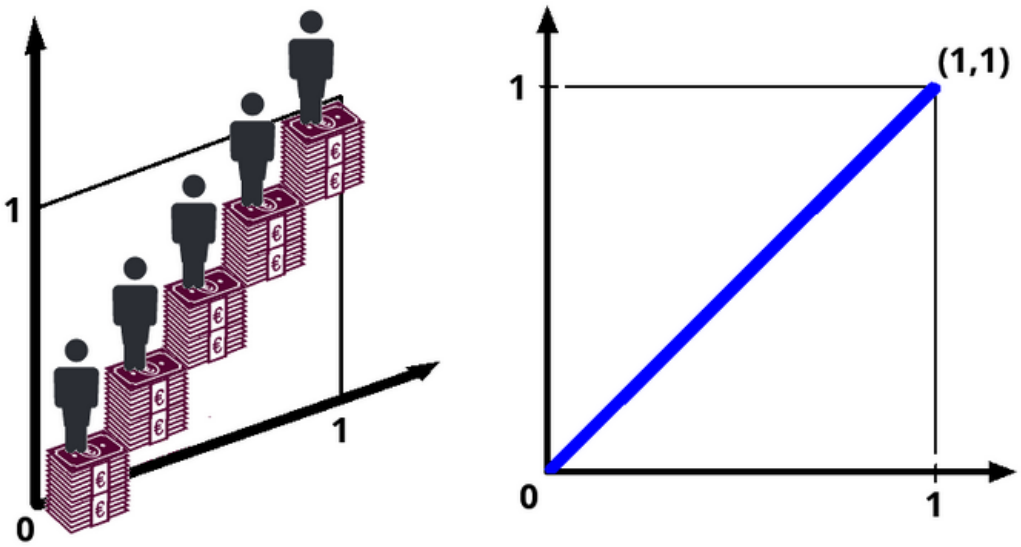


Figure 19: Courbe de Lorenz : répartition régulière

Dans ce cas, la répartition est la plus égalitaire possible. L'escalier se présente comme ceci (à gauche) :

- Et si une seule personne concentre en sa possession l'ensemble de la richesse ? Nous sommes dans l'extrême inverse du cas précédent. Ici, la répartition est la plus inégalitaire possible :

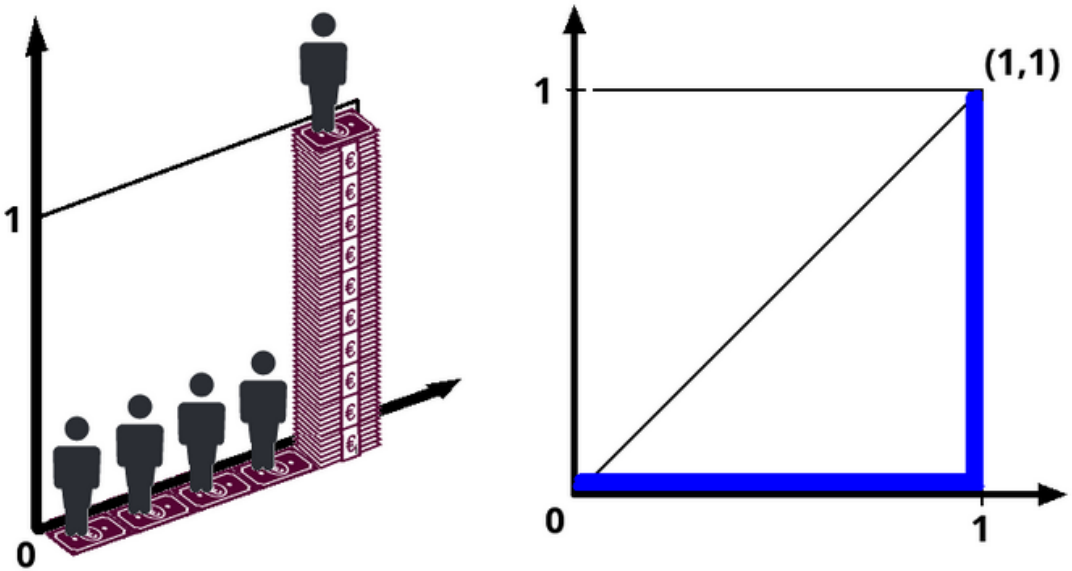


Figure 20: Courbe de Lorenz : répartition inégalitaire

La courbe de Lorenz n'est pas une statistique, c'est une courbe ! Du coup, on a créé l'indice de Gini, qui résume la courbe de Lorenz.

- Il mesure l'aire présente entre la première bissectrice et la courbe de Lorenz. Plus précisément, si on note S cette aire, alors : $gini = 2S$
- L'indice de Gini n'est pas très parlant pour le public. Une autre manière d'exprimer les inégalités plus intelligible est de dire :
  - Les X % les plus riches possèdent Y % de la richesse mondiale, ou bien
  - Les X % les plus riches possèdent autant que les Y % les plus pauvres.

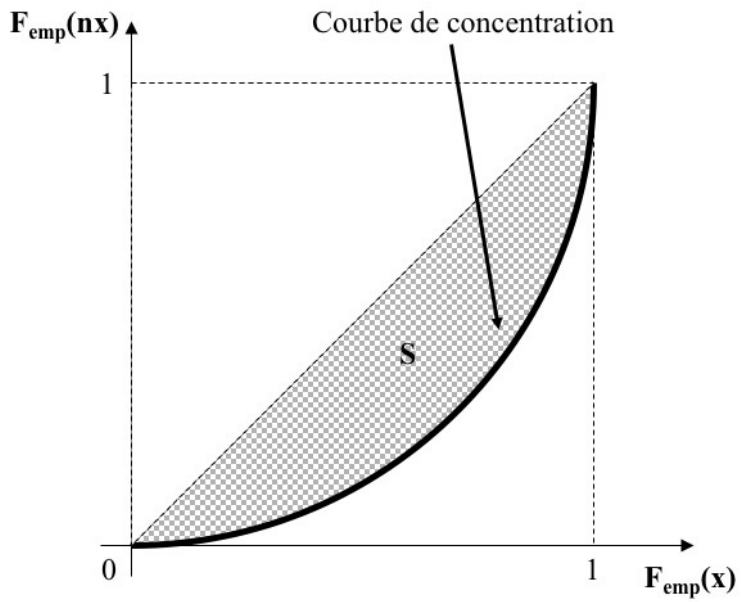


Figure 21: Indice de concentration de Gini

Analyse bivariée

Petit exemple 1/2

- Pour faire un système de recommandation d'album musicaux, vous décidez de faire une étude préalable des données de navigation du site web de téléchargement.
- Les visites des albums du site web permettent de noter les niveaux d'intérêt de 0 à 10. Pour une visite longue et achat : 10!. Pour une première visite d'un nouvel album: 5! Pour une visite courte sans achat : 0!. L'âge et le niveau d'intérêt constituent les variables de l'échantillon
- Les histogrammes suivants concernent un nouvel album. Ils sont bien répartis, mais aucune interprétation âge/intérêt ne peut être évoqué.

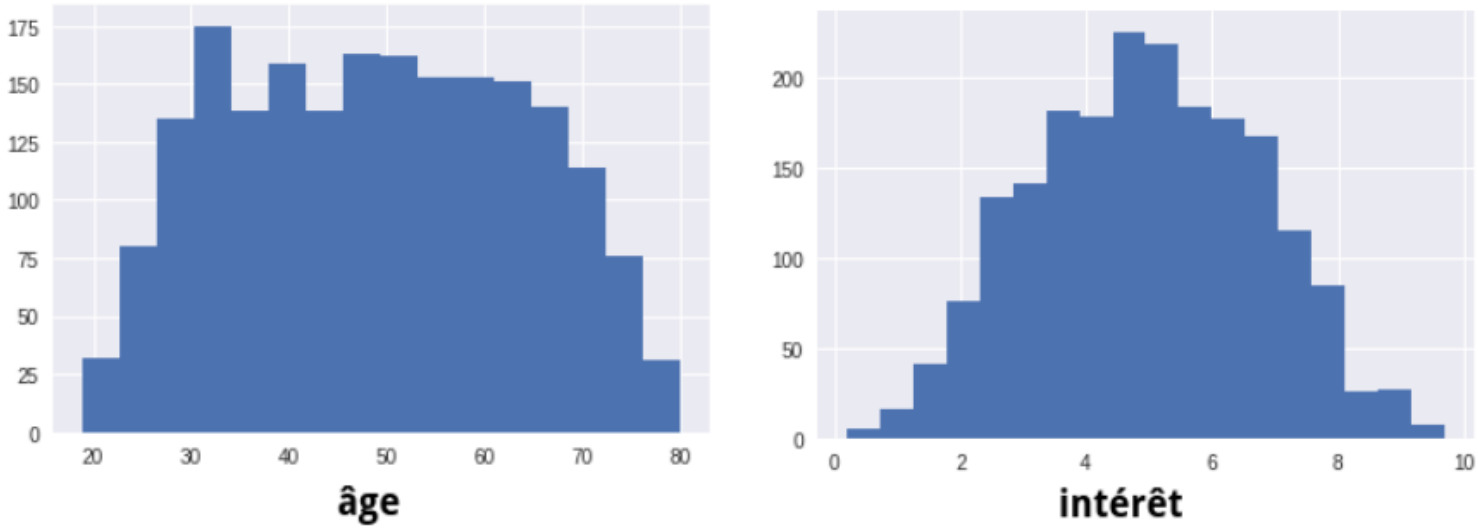
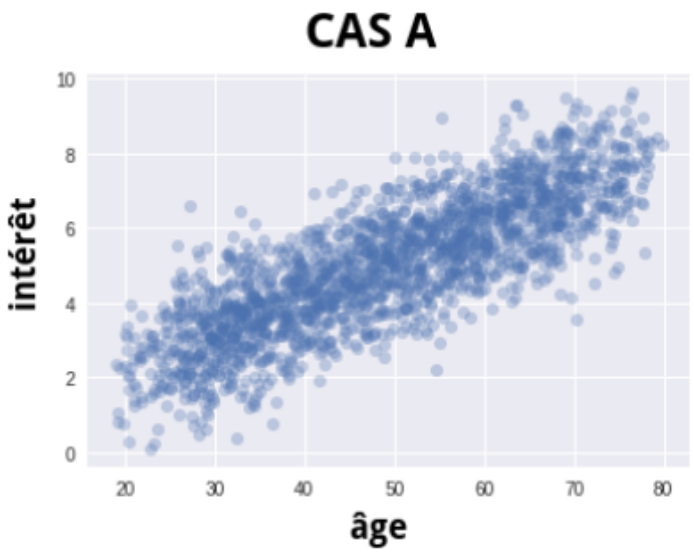
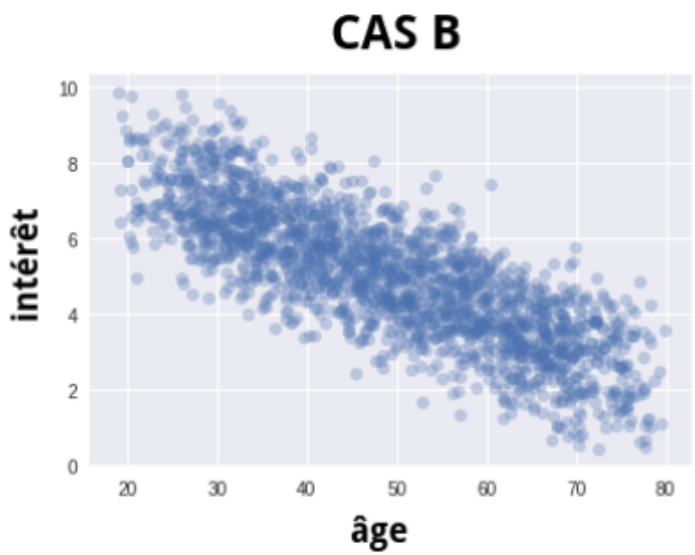


Figure 22: histogrammes

Petit exemple 2/2

- Maintenant, plaçons sur un graphique en 2 dimensions les individus de notre échantillon.
- Sur un graphique de type diagramme de dispersion (scatter plot) on place les coordonnées des individus *niveau d'intérêt=f(âge)*.
- Un point en haut à droite représente une personne plutôt âgée très intéressée par le nouvel album. Au contraire, un point en bas à gauche représente une personne jeune n'aimant pas l'album.
- Cas A : Beaucoup de personnes âgées aiment ce nouvel album
- Cas B : Beaucoup de jeunes aiment cet album !



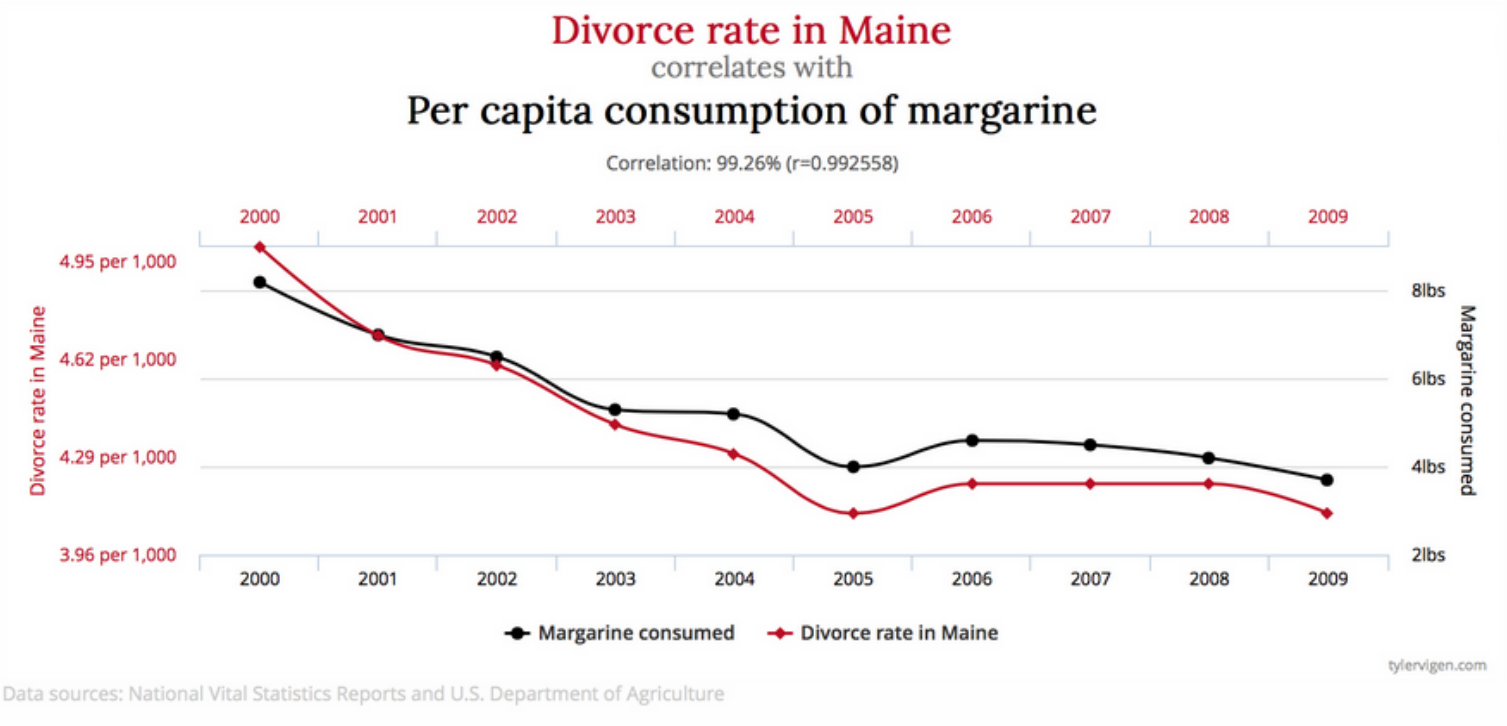


Analyse bivariée

Chercher les corrélations

Dire que deux variables sont **corrélées** signifie que si on connaît la valeur d'une variable, alors il est possible d'avoir une indication (plus ou moins précise) sur la valeur d'une autre variable.

- **Avertissement** : il y a une erreur à ne JAMAIS commettre : dire qu'il y a un lien de cause à effet d'une variable sur l'autre ! (voir le paradoxe de Simpson)
- Si une corrélation existe entre deux variables A et B , est-ce A la cause de B, ou B la cause de A ? Souvent impossible à savoir sans expérimentation. Le plus souvent, c'est un troisième (ou plusieurs) facteur C (d'ailleurs pas toujours observé) qui est la cause de A et de B.



Vous réalisez une enquête dans différents cafés de la ville pour observer discrètement un échantillon de 100 clients et noter la boisson qu'ils ont commandé. On créer deux variables **qualitatives** : X=nom café et Y=boisson préférée. Le tableau de contingence résume la distribution obtenue :

- Chaque valeur du tableau (hors les colonnes Total) est appelé **effectif conjoint** n_{ij}
- L'ensemble des effectifs conjoints est la **distribution conjointe empirique**
- La dernière ligne (TOTAL) est appelée **distribution marginale empirique** de la variable *boisson préférée*, et la dernière colonne (TOTAL) est appelée **distribution marginale empirique** de la variable *nom café*.
- L'ensemble des effectifs conjoints de la première ligne (Chez Luc) est appelée **distribution conditionnelle empirique** de *boisson préférée* étant donné que *nom café = Chez Luc*.

	Café	Thé	Autre	TOTAL
Chez Luc	1	9	0	**10**
Au café Dembas	9	6	5	**20**
Au café Ducoing	20	10	10	**40**
Chez Sarah	20	5	5	**30**
TOTAL	**50**	**30**	**20**	**100**

		Y			Σ
		b_1	...	b_ℓ	
X	a_1	n_{11}	...	$n_{1\ell}$	$n_{1\cdot}$
	⋮	⋮		⋮	⋮
	a_k	n_{k1}	...	$n_{k\ell}$	$n_{k\cdot}$
Σ		$n_{\cdot 1}$	...	$n_{\cdot\ell}$	n

Figure 23: Tableau de contingence

Sur notre relevé de banque, chercher une corrélation entre les variables *montant* et *solde_avt_operation* revient à dire : "Sachant que le solde de votre compte est petit, peut-on s'attendre à ce que le montant de l'opération soit lui aussi petit ?" (ou l'inverse). :

- A priori pas de corrélation entre ces deux variables, car les points sont assez dispersés.

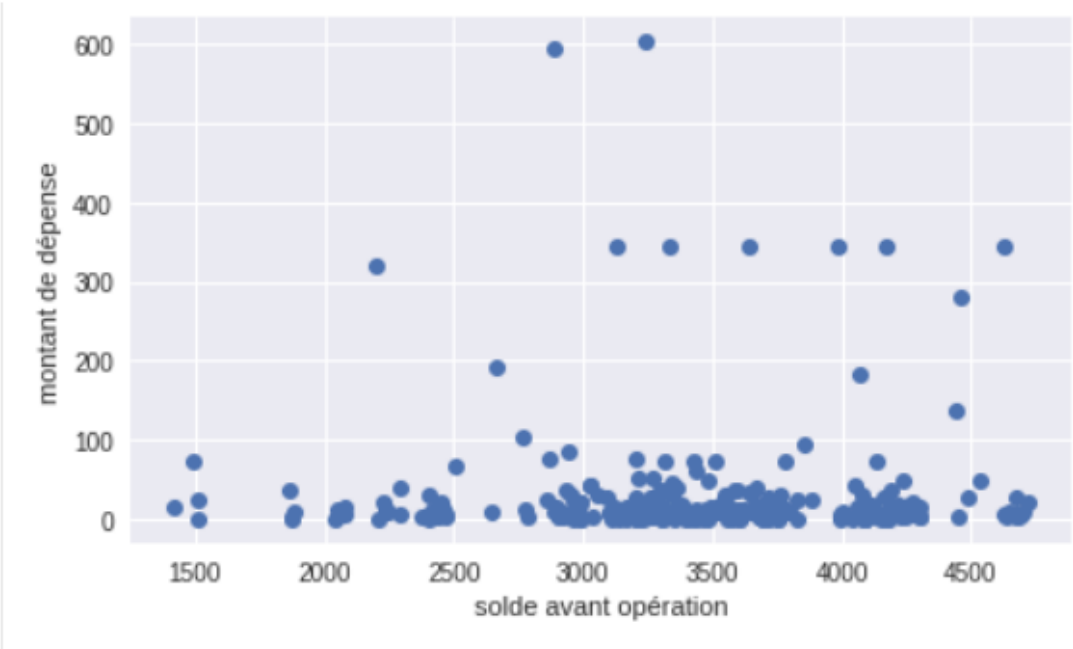


Figure 24: scatter plot de X et Y

- On exprime la **covariance** comme la moyenne des produits des écarts à la moyenne de chaque variable

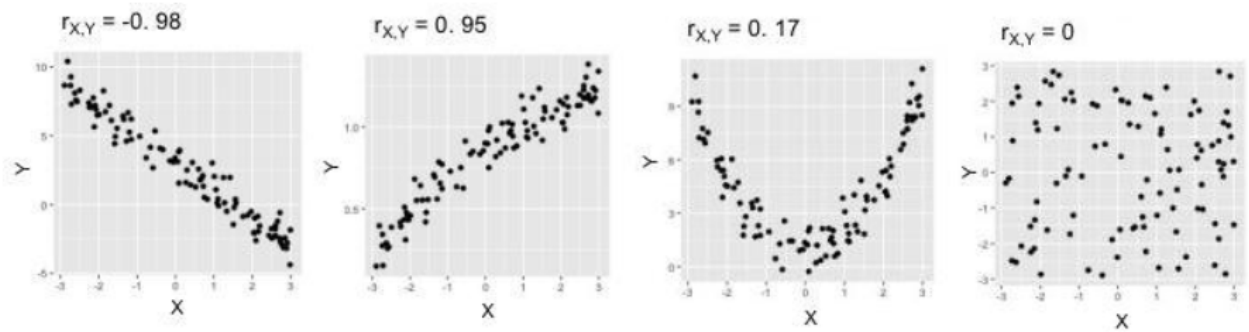
$$s_{X,Y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- Chaque produit des écarts à la moyenne de même signe conforte la corrélation. Chaque produit de signe différent conforte la corrélation inverse.
- La covariance empirique de X et X est la **variance empirique**.

$$s_{X,Y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- La normalisation de la covariance empirique par le produit des écarts types est le **coefficient de corrélation (linéaire, de Pearson)**

$$r_{X,Y} = \frac{s_{X,Y}}{s_X s_Y}$$



- Symétrie :

$$s_{X,Y} = s_{Y,X}$$

- Bilinéarité : soit Z une variable construite à partir de deux autres variables U, V telle que

$$Z = aU + bV$$

alors

$$s_{X,Z} = as_{X,U} + bs_{X,V}$$

Analyse bivariée par régression linéaire

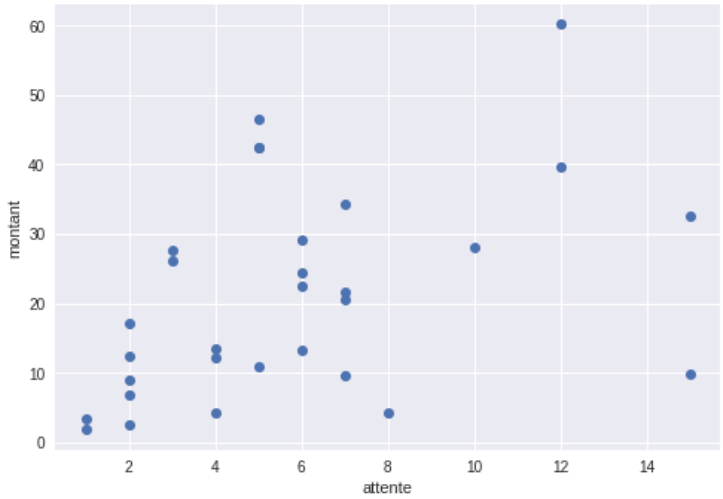
Position du problème

Revenons à notre exemple de compte bancaire. Une variable *attente* (entre deux opérations de courses) a été créée dans le fichier "operations_enrichies.csv". On veut vérifier que le *montant* du ticket de courses est affine par rapport à la durée *attente*. On suppose qu'à chaque course, on achète des produits alimentaires, du stock long terme (boîtes de conserve, surgelé, ..) et des produits non consommables.

- Appelons a le prix moyen des produits par jour, b le prix moyen des produits stockés ou non consommables, y le montant du ticket de caisse.
- Ceci se modélise par la formule (ε représente l'erreur que l'on ne peut annuler avec ce modèle):

$$y = ax + b + \varepsilon$$

- Quel sera le prix de notre prochain ticket de caisse ?



Si on fait varier a et b , on déplace la droite sur le graphique. Minimiser l'erreur revient en fait à placer la droite dans l'alignement général des points. Voici une illustration très pédagogique, les points étant presque dans le même alignement :

- Il existe plusieurs manières de minimiser une erreur. La plus utilisée minimise la somme des carrés de l'erreur $\varepsilon = y_i - \hat{y}_i$.
- On l'appelle méthode des moindres carrés ordinaire (MCO)
- Voici les formules qui permettent d'estimer $\hat{a}$ et $\hat{b}$: $\hat{a} = \frac{S_{X,Y}}{S_X^2}$, $\hat{b} = \bar{y} - \hat{a}\bar{x}$

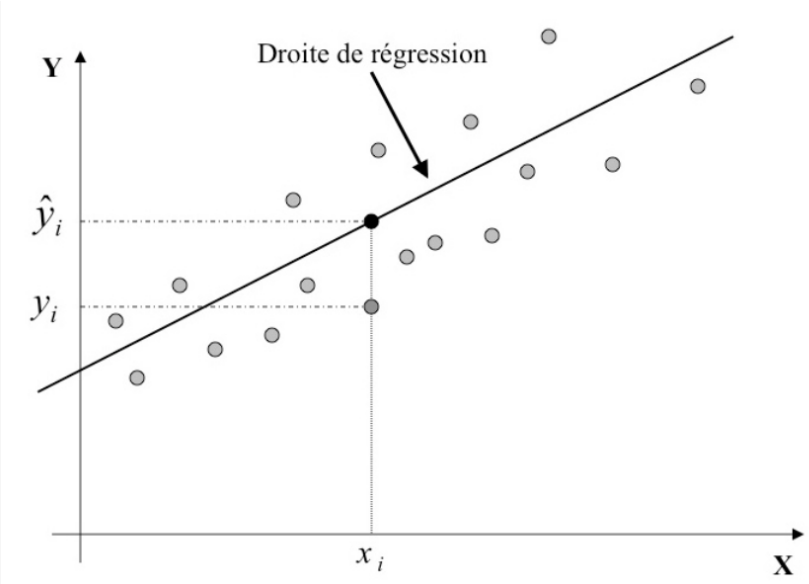


Figure 27: scatter plot montant et attente

Analyse bivariée par régression linéaire

Pour notre exemple et en utilisant les formules précédentes, on obtient les estimations suivantes : $\hat{a} = 1.74$ et $\hat{b} = 10.94$ (droite verte)

- à y regarder de plus près on voit qu'il y a 2 points qui "sortent du lot", on les appelle des outliers. On les écarte car on sait que je ne fais jamais les courses à plus de 15 jours d'intervalle. Ces deux points, pour lesquels attente = 15 jours, correspondent en fait à des retours de vacances (durant lesquels je n'ai pas fait de courses)
- j'obtiens ces nouvelles estimations : $\hat{a} = 3.03$ et $\hat{b} = 5.41$ (droite rouge)
- Avec seulement 2 individus écartés, les résultats changent beaucoup. On dit donc que la régression linéaire (avec estimation par la méthode des moindres carrés) est peu robuste aux outliers.

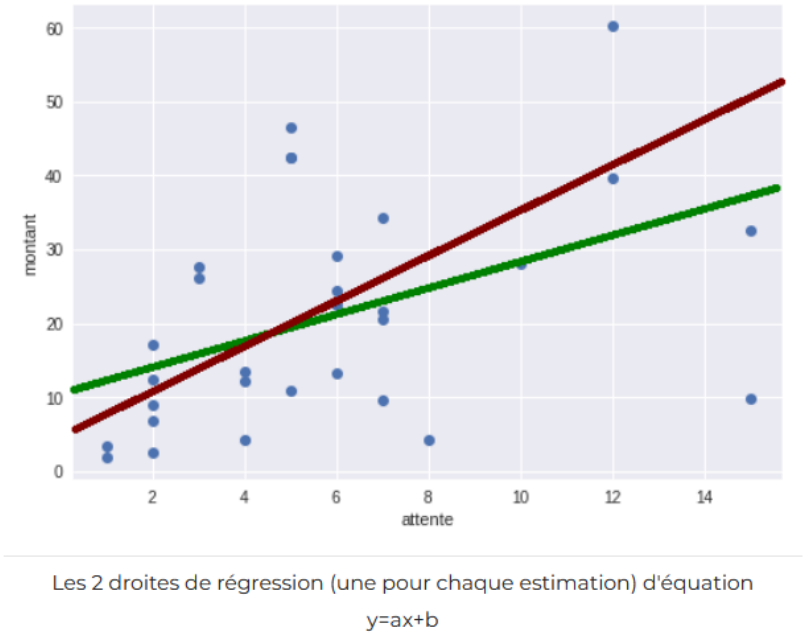


Figure 28: Droite de régression $y = \hat{a}x + \hat{b}$

- Avec le modèle de régression linéaire y = ax+b nous avons cherché à minimiser l'erreur de modèle en minimisant les variations des valeurs de montant autour de la droite de régression.
- Si on avait trouvé un modèle parfait, il n'y aurait plus d'erreur, et donc plus de variation entre les valeurs prédites et les valeurs réelles. On dirait que le modèle a réussi à expliquer la totalité des variations. Les variations autour de la moyenne sont mesurées par la variance. Un modèle parfait aurait expliqué 100 % de la variation. Ce pourcentage est calculé grâce à la formule de décomposition de la variance (analysis of variance, en anglais : ANOVA).
- $SCT = SCE + SCR$, c'est à dire,

$$\sum_{j=1}^{n}(y_j - \bar{y})^2 = \sum_{j=1}^{n}(\hat{y}_j - \bar{y})^2 + \sum_{j=1}^{n}(y_j - \hat{y}_j)^2$$

  - SCT (La Somme des Carrés Totale) traduit la variation totale expliquée,
  - SCE (Somme des Carrés Expliqués) traduit la variation expliquée par le modèle,
  - SCR (Somme des Carrés Résiduelle) traduit la variation inexpliquée par le modèle
- Pour la régression linéaire, le pourcentage de variation expliquée est donné par le coefficient de détermination noté R^2 :

$$R^2 = \frac{SCE}{SCT} = r^2_{X,Y}$$

Analyse bivariée par régression linéaire

Calculs avec python

- Pour calculer le coefficient de Pearson et la covariance, 2 lignes suffisent !
  - *import scipy.stats as st*
  - *import numpy as np*
  - *print(st.pearsonr(depenses["solde_avt_ope"],-depenses["montant"])[0])*
  - *print(np.cov(depenses["solde_avt_ope"],-depenses["montant"],ddof=0)[1,0])*
- Le coefficient de corrélation linéaire se calcule grâce à la méthode *st.pearsonr*. On lui donne ensuite les 2 variables à étudier.
- Rem : on préfère ramener les dépenses en montants positifs, d'où le signe - devant depenses["montant"].
- Un couple de valeurs est renvoyé, le coefficient de corrélation est le premier élément de ce couple, d'où le [0] à la fin de la ligne 4.
- La méthode np.cov renvoie la matrice de covariance, que vous n'avez pas à connaître à ce niveau. Cette matrice est en fait un tableau, et dans ce dernier, c'est la valeur située sur la 2e ligne à la 1e colonne, d'où le [1,0].

Analyse bivariée d'une variable quali et quanti par ANOVA

Position du problème

Passons à l'étude de deux variables dont l'une est qualitative, l'autre quantitative. Nous souhaitons étudier par exemple :

- Les dépenses que vous faites le week-end sont-elles plus grosses qu'en semaine ? (variables montant et weekend) ?
- Le montant d'une opération est-il différent d'une catégorie de dépense à l'autre ? (montant et categ)?
- Vos paiements en carte bancaire sont-ils toujours petits et vos virements importants ? (type et montant)

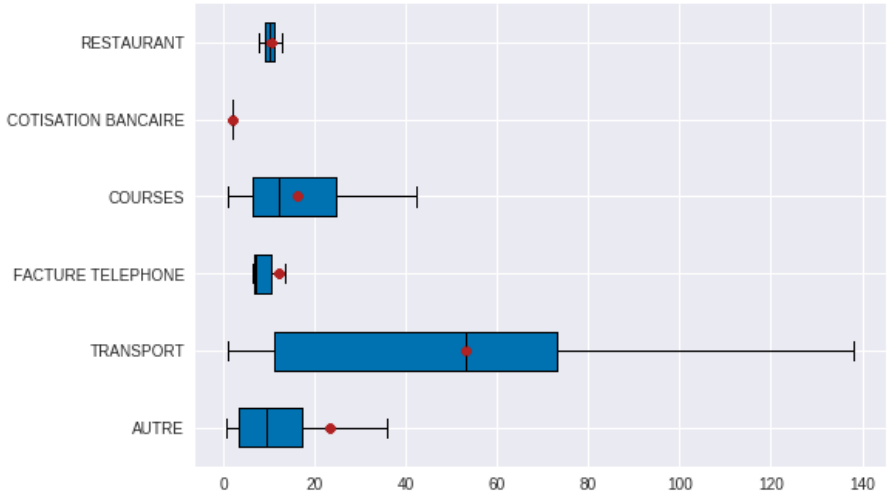


Figure 29: Boîtes à moustaches *Categ =f(montant)*

On pourrait reprendre la formule de la régression linéaire précédente (Y=aX+b), sauf qu'elle implique de multiplier X par a . Or cette fois-ci, X est qualitative, comme notre variable categ. Multiplier une variable qualitative par un nombre n'a aucun sens...

- On fait la supposition que les opérations bancaires ont un montant de référence en commun appelé μ . Ensuite, on considère que le montant de l'opération s'ajuste en fonction de la catégorie *i* de dépense (loyer, transport, courses, etc.). Si une catégorie a des montants qui sont en général inférieurs à μ , alors cet ajustement α_i sera négatif. Dans le cas contraire, il sera positif (par exemple α_{loyer}). On ajoute la contrainte que la somme de tous les α_i soit égale à 0. On sait que l'on commet une erreur car tous les montants ne sont pas les mêmes au sein d'une même catégorie.
- Comme pour le modèle de la régression linéaire, on aura ici aussi un terme d'erreur ε: $Y = \alpha_i + \mu + \varepsilon$
- Le montant μ est estimé par la moyenne de tous les montants, On l'appelle $\hat{\mu}$.
- Pour une catégorie i, α_i est estimé en calculant l'écart entre $\hat{\mu}$ et la moyenne $\bar{y}_i$ des montants de la catégorie *i* , c'est-à-dire $\hat{\alpha}_i = \bar{y}_i - \hat{\mu}$
- Ce modèle est très utilisé en statistiques inférentielles et est appelé analyse de la variance, en anglais **ANalysis Of VAriance** (ANOVA)

Analyse bivariée d'une variable quali et quanti par ANOVA

Évaluation du modèle

Notre modèle est-il de qualité ? Prévoit-on correctement les montants des opérations uniquement à partir de leur catégorie ?

- On espère que notre modèle parvienne à expliquer un gros pourcentage des variations des données. Si c'est le cas, cela signifie que les variables categ et montant sont fortement corrélées.
- La formule utilisée est exactement la même que celle du chapitre précédent : $SCT = SCE + SCR$
- Il faut cependant adapter les notations en introduisant les effectifs n_i des classes i qui sont au nombre de k.
- $SCT = \sum_{j=1}^{n}(y_j - \bar{y})^2$ devient **la variation totale** $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$
- $SCE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ devient **la variation interclasse** $\sum_{i=1}^{k} n_i(\hat{y}_i - \bar{y})^2$
- $SCR = \sum_{j=1}^{n}(y_j - \hat{y}_j)^2$ devient **la variation intraclasse** $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_i)^2 = \sum_{i=1}^{k} n_i s_i^2$
- Le rapport de corrélation, compris entre 0 et 1, donné par :$\eta^2_{Y/X} = \frac{V_{interclasses}}{V_{totale}}$
- **Si** $\eta^2_{Y/X} = 0$ les moyennes par classes sont toutes égales. Il n'y a donc pas a priori de relation entre les variables Y et X.
- Au contraire, **si** $\eta^2_{Y/X} = 1$, alors les moyennes par classes sont très différentes, chacune des classes étant constituée de valeurs identiques : il existe donc a priori une relation entre les variables Y et X.

La méthode d'analyse sera la même pour répondre à toutes les questions suivantes. La seule chose qui change, ce sont les 2 variables étudiées :

- Avez-vous les mêmes catégories de dépenses le weekend et en semaine ? (variables *categ* et *weekend*)
- Vos dépenses sont-elles plus grandes en début de mois qu'en fin de mois ? (*tranche_depense* et *quart_mois*)
- Vos paiements en carte bancaire sont-ils toujours petits et vos virements importants ? (*type* et *tranche_depense*) Y a-t-il des catégories d'opérations qui arrivent toujours au même moment du mois, comme votre loyer par exemple ? (*categ* et *quart_mois*) etc ...

Pour répondre à ces questions, vous pouvez afficher le tableau de contingence comme ceci :

categ	AUTRE	COTISATION BANCAIRE	COURSES	FACTURE TELEPHONE	LOYER	RESTAURANT	TRANSPORT	Total
quart_mois								
1	55.0	7.0	6.0	6.0	NaN	3.0	9.0	86.0
2	45.0	NaN	11.0	1.0	6.0	7.0	6.0	76.0
3	60.0	NaN	7.0	NaN	NaN	4.0	4.0	75.0
4	52.0	NaN	15.0	NaN	NaN	2.0	2.0	71.0
total	212.0	7.0	39.0	7.0	6.0	16.0	21.0	308.0

Résultat du code ci-dessus : le tableau de contingence

Figure 30: Tableau de contingence

- Rappel : Si deux événements I et J sont indépendants, alors on s'attend à ce que le nombre d'individus qui satisfont à la fois I et J (le n_{ij}) soit égal à $f_i \times n_j$
- Au contraire, plus nij sera différent de $f_i \times n_j$, plus I et J sont dépendants.
- Étudier une corrélation entre deux variables qualitatives revient donc à comparer les n_{ij} avec les $f_i \times n_j$. Les n_{ij} sont les nombres dans le tableau de contingence (en dehors des 2 lignes et colonnes TOTAL). On peut créer un autre tableau qui a la même forme que le tableau de contingence, mais qui contient les $f_i \times n_j$

	Café	Thé	Autre	TOTAL
Chez Luc	1	9	0	**10**
Au café Dembas	9	6	5	**20**
Au café Ducoing	20	10	10	**40**
Chez Sarah	20	5	5	**30**
TOTAL	**50**	**30**	**20**	**100**

	Café	Thé	Autre	TOTAL
Chez Luc	5	3	2	**10**
Au café Dembas	10	6	4	**20**
Au café Ducoing	20	12	8	**40**
Chez Sarah	15	9	6	**30**
TOTAL	**50**	**30**	**20**	**100**

Figure 31: Tableaux de contingence observé (à gauche) et simulé (hypothèse d'indépendance) (à droite)

On cherche un indice pour comparer les deux tableaux !

- Pour chaque case i, j, on veut mesurer l'écart quadratique $(n_{ij} - f_i \times n_j)^2$ normalisé par la valeur de $f_i \times n_j$.
- Sachant que $f_i = \frac{n_i}{n}$, la mesure pour une case i, j donne :

$$\xi_{ij} = \frac{(n_{ij} - \frac{n_i}{n} \times n_j)^2}{\frac{n_i}{n} \times n_j}$$

- Plus ξ_{ij} est élevé, moins l'hypothèse d'indépendance est valide pour le cas i, j !
- Si on somme toutes les mesures du tableau (ligne 1 à k) et colonnes (1 à l), on obtient : $\xi_n = \sum_{i=1}^{k} \sum_{j=1}^{l} \xi_{ij}$
- Au delà d'un seuil, appelé p-value (dépendant du risque de se tromper d'hypothèse), on dira que les deux variables sont corrélées. Le test avec le seuil s'appelle le **test du Khi-2** à N-1 degrés de liberté (N étant le nombre d'éléments de la table de contingence)

- En regardant les 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, etc.

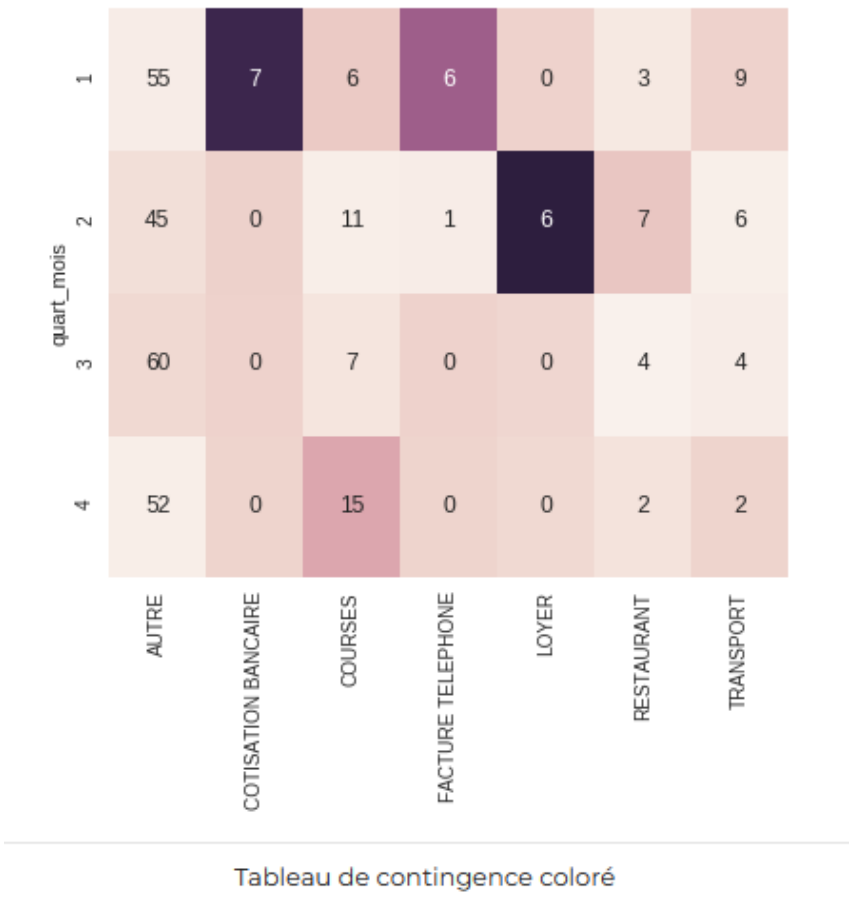


Figure 32: Tableaux de contingence coloré

Analyse exploratoire multidimensionnelle

Pour comprendre l'intérêt, supposons l'exemple d'un échantillon décrivant 2 variables quantitatives x et y. Peu importe ce qu'elles représentent.

- Si l'on commence par une analyse mono-dimensionnelle (ou univariée)

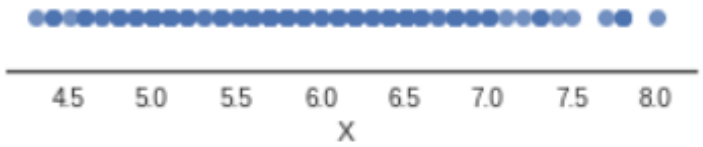


Figure 33: Nuage des points pour le critère *x*

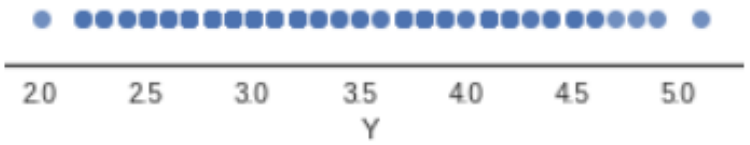


Figure 34: Nuage des points pour le critère *y*

Autres représentations de l'analyse monovariée.

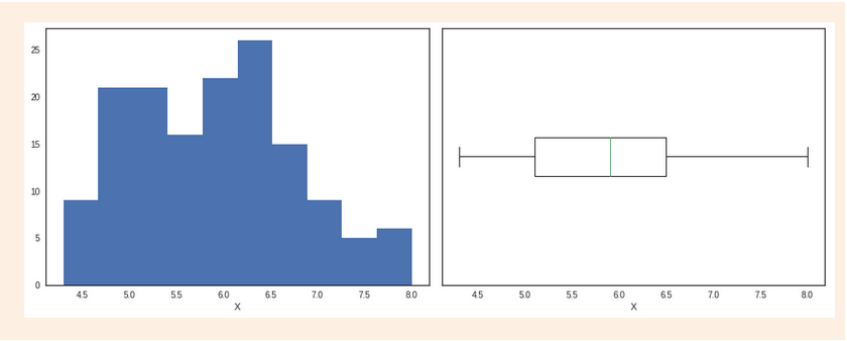


Figure 35: Histogramme et boîte à moustaches du critère x

Représentation bivariée par nuages de points.

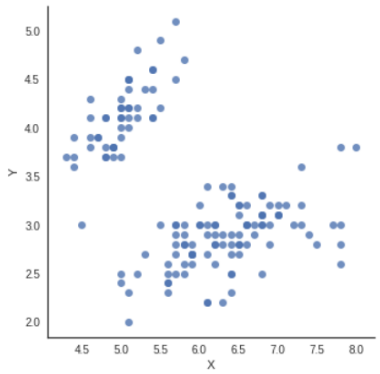


Figure 36: Graphique de dispersion bivariée x, y. On peut voir que les individus sont séparés en 2 groupes bien distincts, pas visible en analyse univariée !

Un deuxième exemple : questionnaire de satisfaction

- Supposons que nous demandons aux étudiants qui suivent ce cours s'ils sont satisfaits, à travers 5 critères sur lesquels ils doivent se positionner sur une échelle de 1 correspondant à "très insatisfait" et 5 à "très satisfait".
- Voici ces 5 critères :
  - Critère 1 : Clarté du cours écrit
  - Critère 2 : Fluidité de la lecture du cours écrit
  - Critère 3 : Les exemples du cours écrit sont-ils faciles à comprendre ?
  - Critère 4 : Richesse du contenu
  - Critère 5 : Êtes-vous satisfait du rasage du professeur ?

	critère 1	critère 2	critère 3	critère 4	critère 5
individu 1	3	3	3	5	5
ind. 2	2	3	2	5	4
ind. 3	2	3	3	4	5
ind. 4	1	1	1	1	1
ind. 5	5	5	4	3	3
ind. 6	4	5	5	2	3
ind. 7	5	5	5	3	3
ind. 8	1	1	1	1	1
moyenne	**2.875**	**3.25**	**3**	**3**	**3.125**

Figure 37: Tableau (imaginaire) des réponses x

Analyse

- Si on regarde les moyennes de chaque critère, on fait une analyse univariée en étudiant chaque colonne une à une. Cela amène à penser que les étudiants ont globalement un niveau de satisfaction qui tourne autour de 3, quel que soit le critère.
- Ce qui nous intéresse ici, c'est le profil des étudiants : le vecteur constitué de leurs 5 réponses.
- On voit par exemple que 2 étudiants ont répondu "1" à tous les critères (les individus 4 et 8).
  - C'est un phénomène classique : à chaque fois que l'on propose un questionnaire de satisfaction, certaines personnes souhaitant exprimer leur mécontentement répondent "très insatisfait" (ou l'inverse) à toutes les questions sans trop regarder l'intitulé de celles-ci.
  - Ce phénomène n'est pas détectable en analyse univariée, car on ne le voit pas en étudiant la moyenne par colonne, par exemple.
  - Il faut décider ensuite si vous souhaitez garder ces individus "revendicatifs" pour la suite des traitements statistiques.
- Il existe 2 groupes bien distincts de personnes : celles qui sont très satisfaites du contenu, mais moyennement du texte (individus 1, 2 et 3), et celles qui sont au contraire très satisfaites du texte, mais moyennement du contenu (individus 5, 6 et 7).

En résumé, l'analyse multivariée est utile quand on souhaite étudier des profils, c'est-à-dire un ensemble de caractéristiques d'un individu.

Analyse exploratoire multidimensionnelle

Quel est l'intérêt d'une étude multidimensionnelle ?

Si un échantillon est représenté par un tableau à 100 000 lignes et 1 000 colonnes, c'est un peu difficile à analyser ! De plus, certains traitements statistiques effectués par ordinateur seront très longs à exécuter : parfois plusieurs jours, plusieurs mois, ou plus !
Pour l'analyse exploratoire multidimensionnelle, nous allons évoquer deux algorithmes emblématiques issues de 2 familles de méthodes non supervisées:

- les méthodes factorielles ;
- les méthodes de classification , aussi appelées de partitionnement de données (clustering).

Analyse exploratoire multidimensionnelle

Quel est l'intérêt de l'analyse en composantes principales (ACP) ?

- L'analyse en composantes principales (ACP) ou Principal component analysis (PCA en anglais) est la plus connue des méthodes factorielles.
- Elle permet de réduire le nombre de variables en trouvant de nouvelles variables qui en synthétisent plusieurs. Une variable synthétique permet de remplacer plusieurs colonnes du tableau par une seule mais en faisant perdre un peu d'information.
- L'ACP permet d'étudier :
  - La variabilité entre les individus, c'est-à-dire quelles sont les différences et les ressemblances entre les individus ;
  - Les liaisons entre les variables : y a-t-il des groupes de variables très corrélées entre elles qui peuvent être regroupées en de nouvelles variables synthétiques ?

- L'algorithme k-means (en français "K-moyennes") se charge de regrouper des individus similaires, c'est-à-dire qu'il va partitionner l'ensemble des individus. Parfois, il est possible de regrouper 100 000 lignes d'un tableau en 3 groupes assez homogènes pour n'étudier finalement que le profil général de chacun de ces 3 groupes, c'est-à-dire 3 lignes !
- Le clustering a de multiples applications. Exemples :
  - En marketing pour segmenter une base de données de clients. Le fait de former des "groupes" de clients et d'étudier leurs caractéristiques (en termes d'âge, de centres d'intérêt, etc.) permet aux marketeurs de cibler leurs campagnes de marketing.
  - En analyse d'image : lorsque 2 pixels d'une photo sont très similaires en termes de couleur, il est possible de les regrouper en une seule couleur. Ainsi, on réduit de manière optimale le nombre de couleurs d'une image, et on réduit donc son poids
  - On peut citer également la classification des espèces. Par exemple, la classification des espèces animales (l'une des plus célèbres) a été introduite pour la première fois au XVIIIe siècle par Linné, naturaliste suédois.

En statistiques, on distingue les traitements supervisés des traitements non supervisés.

- **La classification non supervisée** consiste en l'organisation d'individus en groupes homogènes. En gros, on détermine des classes que l'on ne connaît pas à l'avance.
- **La classification supervisée** consiste à "ranger" les individus dans des classes connues, prédéfinies préalablement. On rencontre parfois le terme de classement, qui est synonyme de classification supervisée.

- Nous travaillons dans un espace vectoriel avec un nombre fini de dimensions (2, 4, 100, 1000 ou beaucoup plus), où chaque individu est représenté par un vecteur (de dimension = nb de variables).
- **Espace euclidien** : chaque composante du vecteur est un nombre réel et on associe à cet espace vectoriel le produit scalaire.
- La notion de distance : il y a plusieurs types de distances
  - La **distance euclidienne** (= à vol d'oiseau)
  - La **distance de Manhattan** appelée aussi taxi-distance lorsqu'il se déplace d'un nœud du réseau à un autre en utilisant les déplacements horizontaux et verticaux du réseau.
- La notion de **nuage de points**: on représente les individus d'un échantillon par des points dans un espace euclidien.
- La **notion d'inertie** : similaire à celle de la mécanique généralisé à un nuage de N_I points à p dimensions. L'inertie de N_I est la moyenne des carrés des distances $d(M_i, G)$ entre les points M_i et leur centre de gravité G. L'inertie totale est égale à :

$$\frac{1}{n}\sum_{i=1}^{n} d(M_i, G)^2$$

- L'inertie du nuage de points N_I est aussi la somme des variances sur toutes les p dimensions $\frac{1}{n}\sum_{i=1}^{n} d(M_i, G)^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 = \sum_{j=1}^{p} Var[j]$

- Pour visualiser des points dans un espace à p dimensions (avec p>2) sur un plan, la solution est d'effectuer une projection orthogonale.
- La projection mathématique, c'est représenter des points à p dimensions dans un espace plus « petit », c'est-à-dire à q dimensions avec q<p .
- Lorsqu'on projette des points, on perd de l'information.

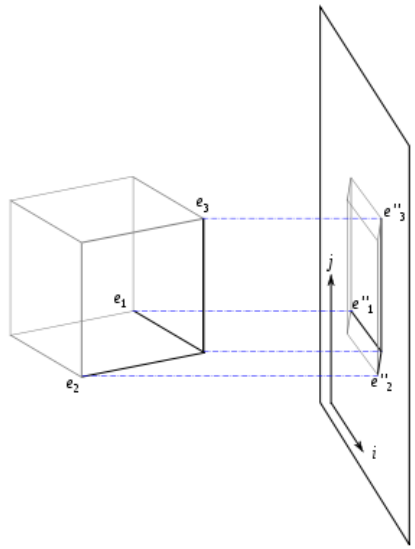


Figure 38: Projection orthogonale d'un cube (à 3 dimensions) sur un plan.

- Y aurait-il des projections qui seraient meilleures que d'autres ? Enjeu : trouver une projection des données qui perde le moins d'informations possibles.
- Pourquoi voit-on mieux un acteur de face que de dessus ? La principale raison est que la forme de l'être humain est allongée : notre hauteur est plus grande que notre largeur. Ainsi, l'image d'un acteur sera plus « allongée » s'il est filmé de face plutôt que du dessus. Son image sera plus « étalée » à l'écran, on a alors plus d'information sur la forme de l'acteur.
- La clé de l'ACP : rechercher la projection pour laquelle l'inertie des points est maximale selon sur un (ou plusieurs) axes.

Vue du dessus

Figure 39: Difficulté de reconnaître une personne vue de dessus par la perte d'information !

- Pour commencer simplement, nous chercherons une projection sur un axe (à 1 dimension) plutôt que sur un plan. Comment positionner cet axe dans l'espace pour que la projection orthogonale du nuage de points soit la plus étalée possible, c.a.d. sur le premier axe principal d'inertie (F1) ?

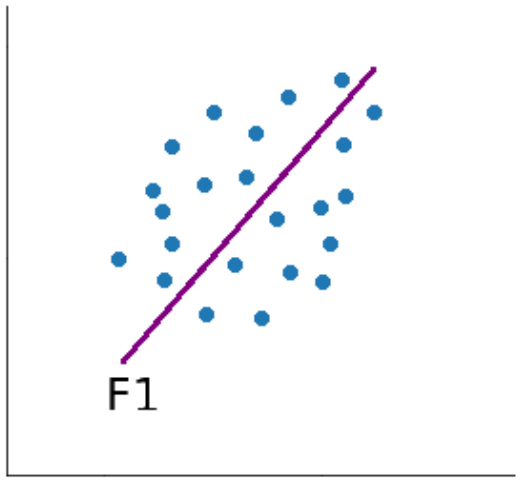


Figure 40: Premier axe F1

- **Second axe d'inertie (F2) ?** Si l'espace des données est à 2 dimensions, alors il n'y a qu'une seule direction possible pour ce second axe : la direction orthogonale au 1er axe d'inertie. Remarque: s'il y a plus de 2 dimensions, il y a alors plusieurs solutions.
- Une fois la direction de ce second axe trouvé, on cherchera le 3e (F3), avec la contrainte qu'il soit orthogonal à tous les précédents.
- Ces axes principaux d'inertie peuvent être vus comme de nouvelles variables calculées à partir des variables déjà existantes (« variables initiales »)

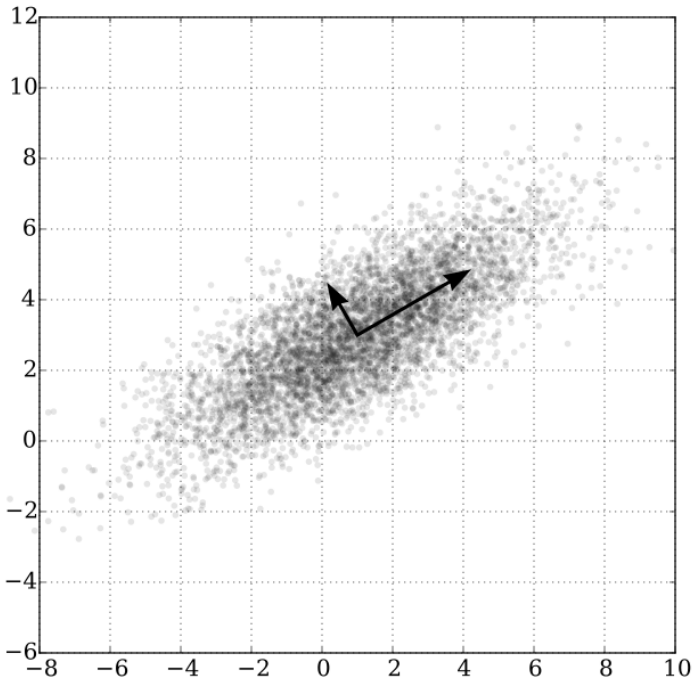


Figure 41: 2 axes principaux d'inertie perpendiculaires à partir de données à 2 dimensions !

- Les axes principaux d'inertie sont des combinaisons linéaires des variables initiales.
- Exemple d'un échantillon décrit par 2 variables x et y. L'axe principal d'inertie peut-être considéré comme une nouvelle variable F1, par exemple de la forme :

$$F1 = 0.8x + 0.6y$$

	x	y	F1
individu 1	1	2	2
individu 2	0	3	1.8
individu 3	-1	1	0.2
...	...	...	...

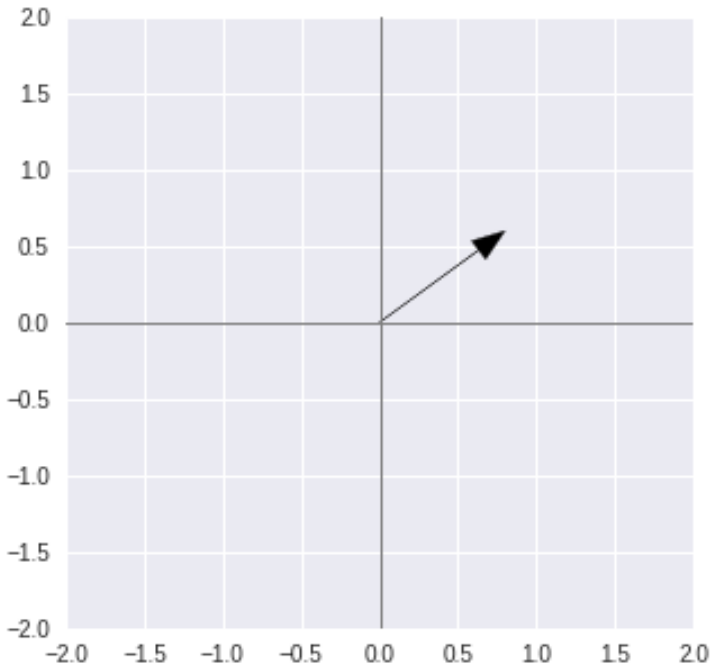


Figure 42: 1ère axe principal en fonction des deux variables x et y

Deux opérations préalables sont nécessaires : la première opération s'appelle le centrage.

- Il est très fréquent que les variables d'un échantillon ne soient pas exprimées dans la même unité parce qu'elles ne représentent pas la même chose.
- Pour comparer des variables qui représentent des quantités différentes : faire en sorte que leurs moyennes soient toutes égales et que leurs variances le soient aussi.
- Classiquement, on s'arrange pour transformer les variables pour que leur moyenne soit égale à 0 et que leur variance soit égale à 1.

- Échantillon 1 :

poids (g)	diamètre (mm)
100	70
95	65

- Échantillon 2 :

poids (g)	diamètre (cm)
100	7
95	6.5

Figure 43: Ces deux échantillons sont identiques, sauf que l'un exprime le diamètre des pommes en millimètres et l'autre en centimètres.

La seconde opération s'appelle la réduction.

- Exemple : les points ci dessous, figure (a), sont disposés selon une forme elliptique. La variance de la variable représentée en abscisses est plus grande que celle de la variable représentée en ordonnées.
- Après avoir centré les données, figure (b) et si on les divise par leur écart-type, on obtient des valeurs dont la variance vaut 1, figure (c).
- Quand on réduit, l'ellipse devient alors un cercle, car la variable en abscisses a maintenant la même importance que celle en ordonnées : il n'y a plus d'aplatissement !
- La formule de centrage-réduction est : $X_{cr} = \frac{X - \bar{X}}{s_X}$, $\bar{X}$ moyenne de X, s_X écart type de X

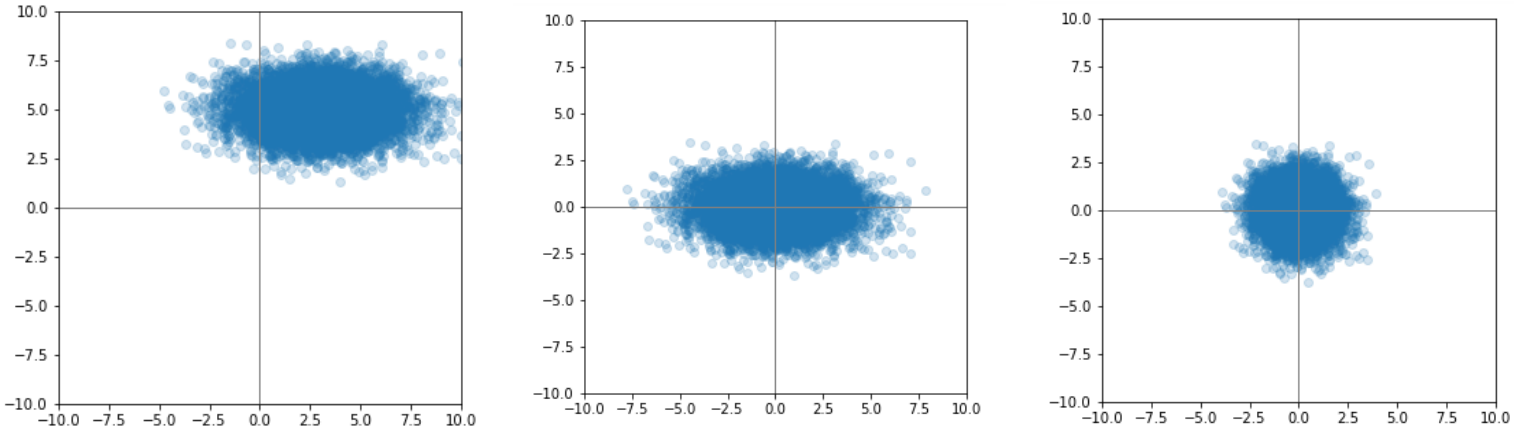


Figure 44: (a) Nuage initial (b) Nuage centré (c) Nuage centré-réduit

Rappelons les 2 objectifs de l'ACP :

- **Objectif 1** :
  - Étudier la variabilité des individus (leurs ressemblances et différences). L'ensemble des points placés dans l'espace R^p forme le nuage des individus, que l'on notera N_I : on recherche les axes d'inertie maximum.
  - Chaque colonne de l'échantillon correspond à une dimension de l'espace $\mathbf{R}^p$ et chaque ligne (chaque individu) correspond à un point dans cet espace.
- **Objectif 2** :
  - Étudier les liaisons entre les variables. Au besoin, regrouper les variables liées en nouvelles variables synthétiques pour réduire le nombre de colonnes (variables).
  - L'espace des individus $\mathbf{R}^n$ a donc autant de dimensions (n) que d'individus et le nuage N_K a autant de points (p) que de variables.

	Dernière mise à jour	Moyenne de classe	Nombre de chapitres	Difficulté
Cours SQL	20 jours	85%	28	2
Cours Python	40 jours	83%	10	3
Cours HTML	60 jours	80%	19	1

Échantillon 3 individus et 4 variables

	Dernière mise à jour	Moyenne de classe	Nombre de chapitres	Difficulté
Cours SQL	-1.225	1.136	1.225	0
Cours Python	0	0.162	-1.225	1.225
Cours HTML	1.225	-1.300	0	-1.225

Table de dessous : centré et réduit

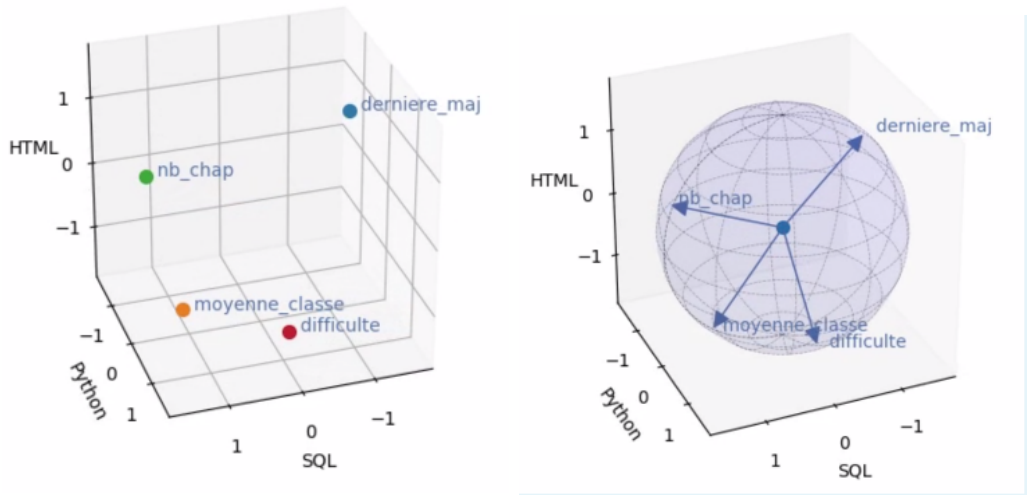


Figure 45: Les variables dans l'espace des individus. (à droite, les flèches représentent mieux les vecteurs)

- Si on calcule l'angle $\widehat{uov}$ (où o est l'origine du repère) entre deux points (correspondant donc à 2 variables u et v), alors cet angle est lié au coefficient de corrélation des variables u et v !
- Il est plus simple visuellement de regarder un angle entre deux flèches.
- **Le cosinus de cet angle est égal au coefficient de corrélation entre u et v**: $cos(\widehat{uov}) = r_{u,v}$ Cette propriété est valable si les données sont centrées, ce qui est toujours le cas en ACP.
- Si les données sont centrées et réduites, les longueurs de toutes les flèches sont égales $|u| = \sqrt{n}$, la racine carrée du nombre d'individus dans notre échantillon.

À ce stade, nous avons 2 espaces totalement différents : $\mathbf{R}^p$ et $\mathbf{R}^n$. Le premier est à p dimensions et contient le nuage d'individus N_I , et l'autre est à n dimensions et contient le nuage des variables N_K .

- Dans chacun des 2 espaces, on a cherché les axes principaux d'inertie.
- Dans $\mathbf{R}^p$, on a vu que l'on pouvait considérer les axes principaux comme de nouvelles variables calculables à partir des variables initiales.
- si l'on place ces nouvelles variables dans $\mathbf{R}^n$, alors celles-ci coïncident exactement avec les axes principaux d'inertie du nuage des variables N_K !
- Autrement dit, étudier les axes d'inertie des individus est équivalent à étudier les axes principaux d'inertie des variables !

	espace	nuage	représentation du nuage	particularité
Objectif 1	$\mathbb{R}^p$	N_I : nuage des individus	par des points	
Objectif 2	$\mathbb{R}^n$	N_K : nuage des variables	par des vecteurs (flèches) partant de l'origine	Toutes les flèches ont la même longueur si les données sont centrées-réduites.

Figure 46:

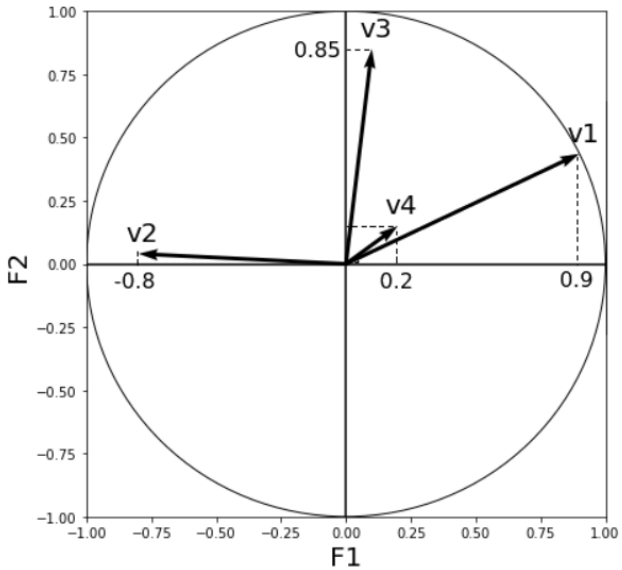
- Notons u_s, un vecteur unitaire portant l'axe de rang s sur lequel la projection du nuage des individus maximise l'inertie, c'est-à-dire la quantité : $\frac{1}{n}\sum_i^{i=n}(H_i^s)^2$, avec la contrainte que u_s soit orthogonal aux $s-1$ directions déjà trouvées (c'est-à-dire u_{s-1},u_{s-2},...,u_1)
- H_i^s est la projection du point i sur u_s dont la norme $\|H_i^s\|$ s'obtient en faisant le produit scalaire entre u_s et i-ème ligne de données.
- On recueille dans le vecteur F_s l'ensemble des normes de H_i^s par le produit $F_s = X.u_s$
- **Maximiser la variance** (c.a.d l'inertie) sur l'axe u_s revient à maximiser $\frac{1}{n}F_s'F_s = \frac{1}{n}u_s'X'Xu_s$ en cherchant le bon vecteur u_s.
- $\frac{1}{n}X'X$ correspond à la **corrélation des variables centrées réduites** (= covariance si les variables sont uniquement centrés)
- Il est montré que le vecteur u_s vérifie $\frac{1}{n}X'Xu_s = \lambda_s u_s$ avec $\lambda_s = \frac{1}{n}F_s'F_s = \frac{1}{n}u_s'X'Xu_s$
- u_s **correspond aux vecteurs propres** et les **valeurs propres** λ_s de la matrice $X'X$ sont les inerties. Pour obtenir λ_s et u_s on réalise **la diagonalisation de la matrice des corrélations** $X'X$

Nous avons 2 espaces :

- l'espace $\mathbf{R}^p$ à p dimensions où l'on a placé le nuage N_I des individus; cela est utile pour déterminer les 2 premiers axes d'inertie (sur un graphique à deux dimensions)
- l'espace $\mathbf{R}^n$ à n dimensions où l'on a placé le nuage N_K des variables, cela est utile pour trouver les corrélations entre variables sur un graphique à deux dimensions : c'est le **cercle des corrélations**

Le graphique montre

- L'axe des abscisses et des ordonnées porte respectivement le premier (F1) et le deuxième (F2) axe d'inertie.
- La projection des flèches (représentant les variables v_i) sur F1 correspondent au coefficient de corrélation entre v_i et F1, compris entre -1 et 1 pour une ACP normalisée.



Étudions donc les corrélations entre les variables initiales et les composantes principales !

- La projection de v_1 sur F_1 vaut 0.9. Cela signifie que $r_{v_1,F1} = 0.9$
- La variable v_2 est corrélée négativement à F1 avec un coefficient de corrélation proche de -1 (ici -0.8) : on dit aussi anticorrélée, c'est-à-dire que, quand v_2 croît, alors F_1 décroît.
- v_3 est très peu corrélée à F_1 , mais l'est fortement à F_2 . V_4 , quant à elle, est très peu corrélée à la fois à F_1 et à F_2.

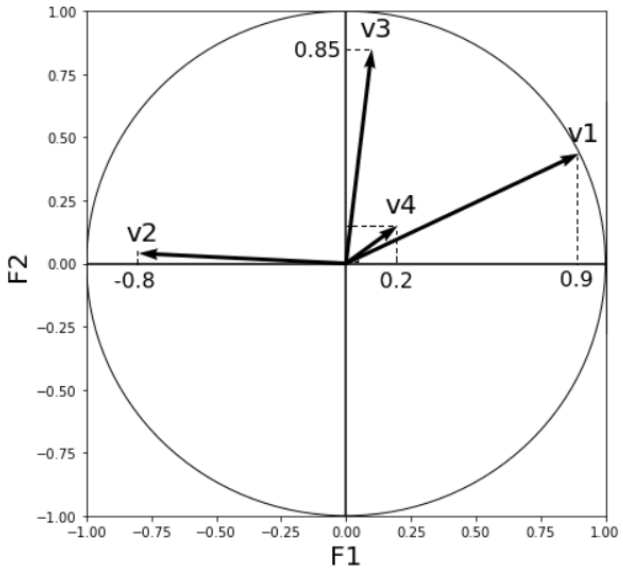


Figure 48: Analyse du cercle de corrélations

Analysons un jeu de données portant sur le cours

- Les variables les plus corrélées à F1 sont : la durée, le nombre de chapitres (nbChapitres), le nombre d'évaluations du cours (nbEvaluations), progression (corr. négative). **Le mode commun qui les unit est la longueur du cours** sur l'axe F1.
- Les variables les plus corrélées à F2 sont : la difficulté (corrélation négative) ; la moyenne de classe ; la proportion de quiz par rapport au nombre total d'évaluations. **On peut interpréter F2 comme la facilité du cours**
- *longueur* et *facilité* résument le jeu de données par ce qui les différencie le plus. On perd de l'information, cependant, **ces 2 variables synthétiques F1 et F2 sont « optimales »**

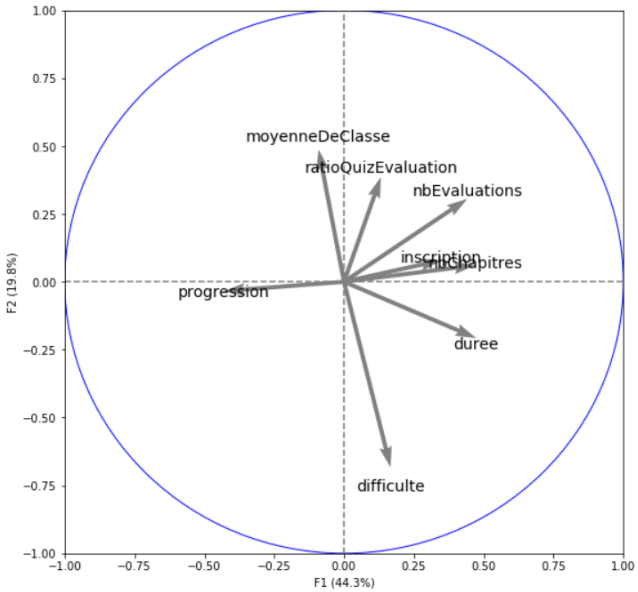


Figure 49: Cercle de corrélations des données des cours

- Les composantes principales peuvent être vues comme de nouvelles variables, c'est-à-dire comme de nouvelles colonnes de notre tableau de données.

Données centrées réduites

```
pd.DataFrame(X_scaled, columns=data.columns, index=data.index)
```

	inscription	progression	moyenneDeC	duree	difficulte	nbChapitres	ratioQuizE	nbEvaluations
titreCours								
Classez_et_segmentez_des_données_visuelles	-1.262416	-0.026684	0.000000	-0.180438	1.687632	-0.749505	0.206125	-0.426401
Initiez-vous_à_la_statistique_inférentielle	-1.104186	-1.011932	-0.909501	-0.501216	-0.112509	0.461817	0.555914	0.426401
Découvrez_les_librairies_Python_pour_la_Data_Science	-1.049414	0.958565	1.765501	-0.715067	-0.112509	-1.112902	-2.592181	-1.279204
Devenez_mentor_sur_OpenClassrooms	-0.976385	2.140863	0.428000	-1.142771	-1.912649	-0.628373	1.605278	-0.426401
Initiez-vous_à_l'algèbre_relationnelle_avec_le_langage_SQL	-0.270436	-0.683516	0.695500	0.354192	-0.112509	0.946345	-0.493451	0.426401

Calcul des composantes principales

```
import pandas as pd
from sklearn import decomposition, preprocessing
X_scaled = preprocessing.StandardScaler().fit_transform(data.values)
pca = decomposition.PCA(n_components=8)
X_projected = pca.fit_transform(X_scaled)
pd.DataFrame(X_projected, index=data.index, columns=["F"+str(i+1) for i in range(8)])
```

	F1	F2	F3	F4	F5	F6	F7	F8
titreCours								
Classez_et_segmentez_des_données_visuelles	-0.746426	-1.310727	0.271104	-1.237802	-1.039614	0.209733	0.479134	0.161148
Initiez-vous_à_la_statistique_inférentielle	0.355009	0.056853	1.047643	-0.958762	0.494815	-0.874848	0.893100	0.290791
Découvrez_les_librairies_Python_pour_la_Data_Science	-2.726654	-0.486699	2.701208	-0.315756	0.314049	0.255670	0.000008	0.153843
Devenez_mentor_sur_OpenClassrooms	-2.432252	2.051605	1.502815	0.137795	0.354128	1.109119	-0.238830	0.150274
Initiez-vous_à_l'algèbre_relationnelle_avec_le_langage_SQL	0.852377	0.341844	0.886664	-0.710740	0.536135	0.013995	0.244542	-0.029600

```
f1 = pca.components_[0]
print(f1)
```

```
[ 0.34933787 -0.4302033  -0.09120044  0.4733088   0.16673833  0.46937238  0.13200943  0.44160032]
```

Figure 50: Nouvelles variables synthétiques du tableau des données des cours

- Les variables synthétiques sont des combinaisons linéaires d'autres variables :

$$\begin{aligned} F_1 = & 0.35 * inscription \\ & -0.43 * progression \\ & -0.09 * moyenneDeClasse \\ & +0.47 * duree \\ & +0.17 * difficulte \\ & +0.47 * nbChapitres \\ & +0.13 * ratioQuizEvaluation \\ & +0.44 * nbEvaluations \end{aligned}$$

Figure 51: Combinaison linéaire de F1 avec les coefficients les plus influents

Remarques :

- Les petites flêches du cercle de corrélation correspond à des variables mal représentées par le plan factoriel 2D, en raison de la projection de l'hypersphère sur le cercle choisi par l'ACP. Il faudrait les regarder sur un autre plan factoriel pour mieux les analyser.
- Il est donc difficile de considérer les correlations des petites flêches car l'analyse est faussée par perte d'information pour les variables dont l'axe est très orthogonal.

Nous nous intéressons ici à l'espace $\mathbf{R}^p$, dans lequel se situe le nuage des individus N_I . Nous souhaitons projeter ce nuage sur le premier plan factoriel, composé des 2 premières composantes principales F1 et F2.

- Il faut interpréter ce graphique (plan factoriel des individus) en parallèle du cercle des corrélations qui indique quelles variables sont très corrélées (ou anticorrélées) à F1 et F2.

Qu'est-ce qui différencie les individus de grande ou petite abscisse ?

Ex: Qu'est-ce qui différencie le cours *Framework Symphony* de *Python pour data sciences* ?

- Les axes principaux d'inertie (F1,F2) synthétisent des variables déjà existantes. Se déplacer dans le sens des abscisses croissantes de F1 (très corrélées aux variables *durée*, *nombre de chapitres* et *nombre d'évaluations*), c'est aller vers un cours à **longueur croissante**.

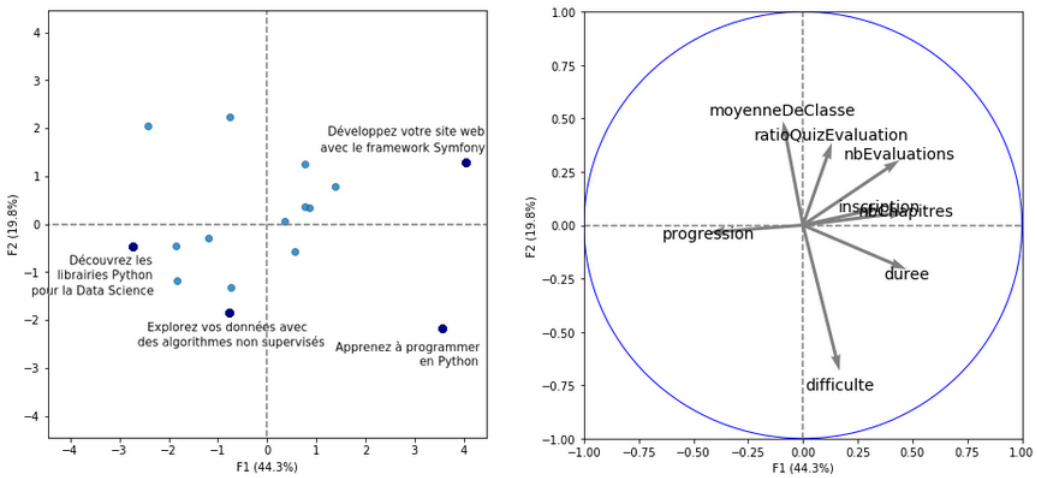


Figure 52: Analyse par les deux plans factoriels

Combien d'axes d'inertie peut-on garder ?

- Dans un espace à p dimensions, il y a au maximum p axes orthogonaux.
- Pour décrire parfaitement n individus, on a donc besoin au maximum de $n-1$ axes. Si $p=3$, pour capter 100 % de l'information, il suffit de faire passer un plan (2D) par ces 3 points. Donc, 2 axes suffisent donc pour 3 points.
- Le nombre maximal de composantes d'une ACP sera le $min(p, n-1)$. Ainsi, un échantillon de 2 000 individus décrits par 1 000 variables conduit à 1 000 composantes.
- On projette les données sur les axes principaux d'inertie, et ceux-ci sont ordonnés selon l'inertie du nuage projeté : de la plus grande à la plus petite.
- Ce diagramme décrit le % d'inertie totale de chaque axe et montre l'éboulis des valeurs propres et le cumul (en rouge) de l'inertie.

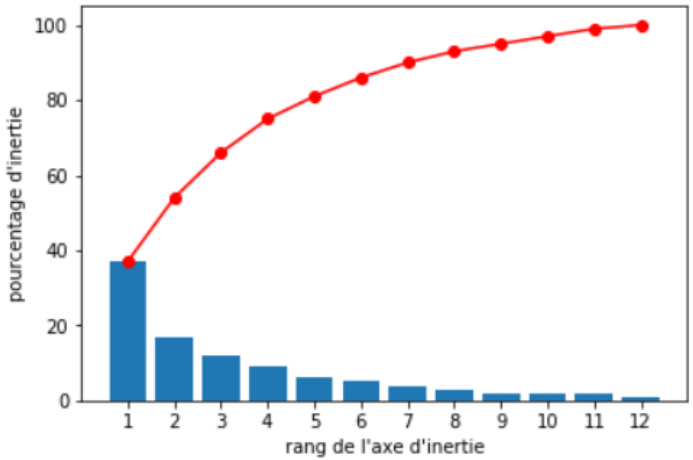


Figure 53: Analyse par les deux plans factoriels

Combien d'axes d'inertie peut-on trouver ?

- Les % d'inertie donnent une information sur la structure des données.
- **Aucune structure** = les variables n'ont aucune corrélation entre elles. L'inertie est équitablement répatie entre les axes (100/p).
- Structure extrême = toutes les variables sont corrélées deux à deux avec un coefficient de corrélation de 1. Il n'y a besoin que d'un seul axe pour capter 100% d'inertie.
- On a tendance à ne **pas considérer comme importants** les axes dont l'**inertie associée est inférieure à (100/p)%** car ils représentent moins de variabilité qu'une seule variable initiale. Ce critère est appelé **critère de Kaiser**.
- La « **méthode du coude** » consiste à repérer l'endroit à partir duquel le pourcentage d'inertie diminue beaucoup plus lentement lorsque l'on parcourt le diagramme des éboulis de gauche à droite.
- Il est **fréquent** de n'analyser **que le premier plan factoriel**.

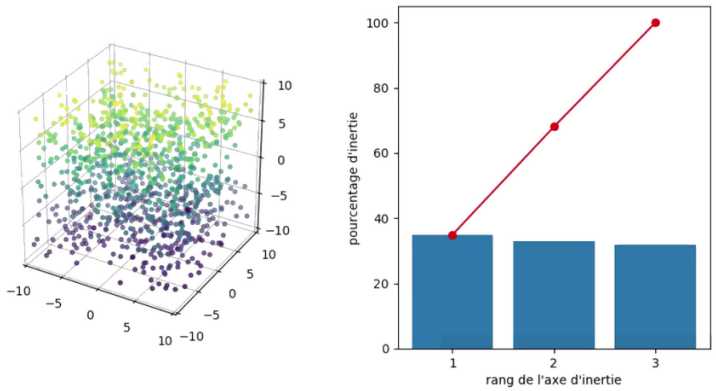


Figure 54: Analyse par les deux plans factoriels

Analyse exploratoire multidimensionnelle

Programme Python PCA sur l'évaluation de différents cours

Nous allons ici analyser les jeux de données que nous avons détaillés dans cette partie.
Ce code est accessible dans le fichier ***TP_cours_ACP.ipynl*** disponible sur moodle.
Prenez le temps de bien lire, de bien comprendre et/ou de jouer avec le code du TP.

Analyse exploratoire multidimensionnelle

Programme Python PCA d'un jeu de données mystères provenant d'un scanner 3D

- Ce code est accessible dans le fichier ***pca_mystery.py*** disponible sur moodle
- Recopier les différentes lignes du fichier .py pour construire un fichier .ipynl
- Dans cet échantillon, chaque individu est un point d'un animal mystère.
- Cet échantillon se rapproche de ce qu'utilisent les personnes qui pratiquent la modélisation 3D (architectes, vidéastes, ingénieurs en mécanique, physiciens, etc.). D'où le parallèle entre nos nuages de points de statisticiens et l'inertie des objets réels qu'étudient les physiciens.
- Quand vous projetez et visualisez vos données sur F1 et F2, vous faites une projection orthogonale en 2D d'un objet en 3D.
- L'ACP a trouvé la meilleure projection : celle qui montre la plus grande inertie.
- Conclusion : Si vous voulez prendre une photo d'un animal, le meilleur angle pour avoir le plus de détails possible sera donc de profil, comme vous pouvez le voir dans cet exemple.

Ce code est accessible dans le fichier ***pca_bag_of_words.py*** disponible sur moodle
Recopier les différentes lignes du fichier .py pour construire un fichier .ipynl

L'ACP est une méthode qui nécessite un peu d'entraînement et pour laquelle il faut être prudent, surtout au début.

- Il est fréquent d'être un peu perdu dans toutes ces flèches, ces points, les plans factoriels, les axes d'inertie, etc. Cela conduit parfois à des interprétations un peu incertaines ou erronées. Mais, vous avez toujours la possibilité de vérifier vos analyses en revenant aux données initiales.
- Si les variables d'un groupe vous semblent corrélées, alors calculez les coefficients de corrélation entre elles (ou la matrice de corrélation, c'est plus rapide)
- Si certains individus vous semblent similaires (car ils ont des abscisses ou des ordonnées à peu près égales sur un plan factoriel), alors vérifiez-le sur vos données initiales. Par exemple, si ces individus ont des abscisses similaires, alors prenez les variables fortement corrélées à F1, et vérifiez si vos individus ont des valeurs semblables pour ces variables.
- Rappelons les 2 objectifs de l'ACP:
  - Étudier la variabilité des individus (leurs ressemblances et leurs différences).
  - Étudier les liaisons entre variables, et trouver de nouvelles variables qui synthétisent les groupes de variables très liées.

- En ACP, nous sommes limités aux corrélations linéaires. Pour passer outre ce problème, on peut utiliser l'ACP avec noyau, ou kernel PCA en anglais.
- Le coefficient r est très sensible aux outliers : c'est donc le cas aussi pour l'ACP.
- Parfois, un axe d'inertie n'était dû qu'à un petit groupe d'individus (ou même à un seul) : ce sont les outliers. Un individu situé très loin de tous les autres a tendance à « attirer » dans sa direction l'un des axes d'inertie (bien souvent le premier). Si l'outlier (ou le groupe d'outliers) ne présente pas d'intérêt dans votre analyse, alors il suffit de ne pas analyser l'axe auquel il contribue fortement à quasiment 100
- il existe d'autres méthodes factorielles permettant de remédier à cela, comme l'Analyse des Correspondances Multiples **(ACM)** pour des variables qualitatives, ou l'Analyse Factorielle des Données Mixtes **(AFDM)**.