TP 24
Statistiques

Ce TP est en même temps un chapitre de mathématiques du premier semestre, nommé *statistiques descriptives*. Le mot *descriptif* signifie qu'on s'intéresse avant tout à différentes façon de représenter des données et d'en extraire des informations, mais qu'on ne prétend pas faire de prédictions avec (*statistiques inférentielles*).
Pour l'illustrer, nous utiliserons d'abord le fichier joint `tips.csv` qui contient des statistiques sur les pourboires laissés au restaurant (en anglais *tip*), en fonction du montant total de la facture, du jour et du moment de la journée. La question générale est : quels facteurs influencent la facture et le pourboire laissés par les clients ? On pourra ensuite utiliser le fichier `diamonds.csv` contenant des données sur les diamants extraits dans une mine (qualité, profondeur, dimensions) et étudier les caractéristiques qui influencent leur prix.

I Introduction

Définition 1. En statistiques, l'ensemble sur lequel on veut étudier les données s'appelle une **population** et chaque élément un **individu**. Les quantités elles-mêmes que l'on étudie s'appelle des **caractères**. La taille de la population s'appelle **effectif**.

La population peut par exemple être toute la population française, ou bien tous les élèves d'une classe (le mot individu a alors son sens habituel). Dans le fichier d'exemples la population est l'ensemble des repas au restaurant qui sont étudiés, chaque repas est considéré comme un individu, dont les caractères sont notamment le prix du repas et le montant du pourboire.

Définition 2. On distingue deux types de caractères :

1. Les caractères **quantitatifs** sont mesurés par un nombre réels (longueur, prix, température, etc). On peut par exemple les exprimer dans différentes unités, s'intéresser à leur somme et à leur moyenne.
2. Les caractères **qualitatifs** sont donnés par une « étiquette » qui peut être n'importe quel mot ou concept (couleur, nationalité, un des sept jours de la semaine, une réponse binaire telle que oui/non, etc). Il n'aurait pas de sens de les additionner ou de calculer leur moyenne, par contre l'opération de base est de regrouper les individus selon une valeur commune du caractère, en fournissant les données sous forme de liste d'effectifs (par exemple combien de personnes de chaque nationalité ; combien de repas au restaurant chaque jour de la semaine ; combien de fumeurs ou non-fumeurs).

Définition 3. Une **série statistique à une variable** est un ensemble de valeurs qu'on note $(x_i)_{i=1,\dots,N}$, où N est l'effectif total, x_i est la valeur du caractère qu'on étudie pour l'individu i. En général l'ordre des individus n'a pas d'importance, on pense donc à la série comme à un « paquet de valeurs » en vrac.

Une série statistique à plusieurs variables sera donnée par la mesure de plusieurs caractères $x_i, y_i, z_i, \dots$ pour une **même** population (donc c'est le même indice $i = 1, \dots, N$).
Les données sont soient présentées comme une simple liste

individu	1	2	$\cdots$	N
valeur	x_1	x_2	$\cdots$	x_N

(1)

mais parfois aussi en **regroupant par valeur** (on dit aussi **modalités**) : on indique pour chaque valeur possible de la série le nombre d'individus ayant cette valeur, attention aux notations car cette fois on s'intéresse avant tout aux valeurs et pas aux individus

valeur	v_1	v_2	$\cdots$	v_q
effectif	n_1	n_2	$\cdots$	n_q

$$\sum_{j=1}^{q} n_j = N \qquad (2)$$

Dans ce cas, la **fréquence** de la valeur v_j est $\boxed{f_j = \dfrac{n_j}{N}}$, un nombre dans $[0, 1]$ souvent exprimé en pourcentage.

II Démarrage en Python

Pour ce TP nous utilisons la bibliothèque Pandas. Elles permet de travailler sur un jeu de données avec une grande flexibilité, en intéragissant naturellement avec Numpy et avec Matplotlib. On peut même retrouver assez facilement les concepts que nous avons vus avec SQL (requête `SELECT` avec `WHERE`, `ORDER BY`, `GROUP BY`), tout dans Python. Cependant l'objectif n'est certainement pas d'approfondir sur le fonctionnement de la bibliothèque Pandas : son gros avantage est de charger d'un seul coup tout un fichier CSV de données, mais nous pourrons toujours les convertir en listes Python au sens le plus usuel possible et écrire nos fonctions sur des listes.
On la charge donc avec les commandes

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

puis on charge d'un seul coup tout le fichier de statistiques avec

```
tips = pd.read_csv("tips.csv")
```

Observez le résultat

```
>>> print(tips)
     total_bill   tip     sex smoker   day    time  size
0         16.99  1.01  Female     No   Sun  Dinner     2
1         10.34  1.66    Male     No   Sun  Dinner     3
2         21.01  3.50    Male     No   Sun  Dinner     3
3         23.68  3.31    Male     No   Sun  Dinner     2
4         24.59  3.61  Female     No   Sun  Dinner     4
..          ...   ...     ...    ...   ...     ...   ...
239       29.03  5.92    Male     No   Sat  Dinner     3
240       27.18  2.00  Female    Yes   Sat  Dinner     2
241       22.67  2.00    Male    Yes   Sat  Dinner     2
242       17.82  1.75    Male     No   Sat  Dinner     2
243       18.78  3.00  Female     No  Thur  Dinner     2

[244 rows x 7 columns]
```

Tout à gauche, les individus sont numérotés à partir de 0 (mais le numéro n'a aucune importance). En haut, les colonnes : facture totale, pourboire, genre (homme/femme), fumeur (oui/non), jour, moment de la journée (Lunch/Dinner), nombre de clients à table. Chacune de ces colonnes correspond à une série statistique.

Exercice 1. Quelles sont les variables quantitatives et les variables qualitatives ici ?

La variable `tips` est de type *Pandas DataFrame*, représente d'un coup tout ce tableau de données. On peut accéder à chacune des colonnes exactement comme avec les listes et les dictionnaires, chaque colonne est de type *Series* (série statistique à une variable).

```
>>> print(tips["total_bill"])
0      16.99
1      10.34
2      21.01
3      23.68
4      24.59
       ...
239    29.03
240    27.18
241    22.67
242    17.82
243    18.78
Name: total_bill, Length: 244, dtype: float64
```

Remarquez le tout dernier mot en bas à droite : chaque série a un type des données, hérités de `numpy`, ici `float64` indique bien que ce sont des nombres à virgule flottante.

```
>>> print(tips["smoker"])
0       No
1       No
2       No
3       No
4       No
      ...
239     No
240    Yes
241    Yes
242     No
243     No
Name: smoker, Length: 244, dtype: object
```

On peut obtenir le résumé de ces informations pour toutes les colonnes :

```
>>> tips.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   total_bill  244 non-null    float64
 1   tip         244 non-null    float64
 2   sex         244 non-null    object
 3   smoker      244 non-null    object
 4   day         244 non-null    object
 5   time        244 non-null    object
 6   size        244 non-null    int64
dtypes: float64(2), int64(1), object(4)
memory usage: 13.5+ KB
```

Les mots `non-null` qui apparaissent signifient que Pandas ajoute la possibilité (qui n'existe pas vraiment avec Numpy) d'avoir des données manquantes. Comme en SQL, une moyenne ou un compte total ne sont pas les mêmes si on ignore les données manquantes ou si on les complète par zéro.
À tout moment, les colonnes peuvent être converties en banales listes Python :

```
>>> tips["total_bill"].to_list()
[16.99, 10.34, 21.01, 23.68, 24.59, ...]
```

III Divers indicateurs statistiques à une variable

III.1 Les classiques

D'abord les indicateurs suivants sont bien connus.

Définition 4. Soit une série statistique $(x_i)_{i=1,\dots,N}$ à une variable.

1. Le **maximum** de la série est une valeur M de la série telle que $\forall 1 \leqslant i \leqslant N$, $x_i \leqslant M$.
2. Le **minimum** de la série est une valeur m de la série telle que $\forall 1 \leqslant i \leqslant N$, $m \leqslant x_i$.
3. La **moyenne** de la série est la valeur $\overline{x}$ définie par

$$\boxed{\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i} \tag{3}$$

Propriétés 5. 1. La moyenne est toujours entre le minimum et le maximum : $m \leqslant \overline{x} \leqslant M$.

2. Si on multiplie tous les termes de la série par une constante $\alpha \in \mathbb{R}$, alors la moyenne est aussi multipliée par α : $\overline{\alpha x} = \alpha \overline{x}$.
3. Si on additionne un nombre $\beta \in \mathbb{R}$ à tous les termes de la série, alors on ajoute aussi β à la moyenne : $\overline{x + \beta} = \overline{x} + \beta$.

Démonstration. 1. Pour tout $1 \leqslant i \leqslant N$, on écrit $m \leqslant x_i \leqslant M$ et on somme ces inégalités :

$$\sum_{i=1}^{N} m \leqslant \sum_{i=1}^{N} x_i \leqslant \sum_{i=1}^{N} M \tag{4}$$

ce qui donne $N \times m \leqslant \sum_{i=1}^{N} x_i \leqslant N \times M$, puis diviser par N partout.

2. On a

$$\overline{\alpha x} = \frac{1}{N} \sum_{i=1}^{N} \alpha x_i = \frac{1}{N} \alpha \sum_{i=1}^{N} x_i = \alpha \times \frac{1}{N} \sum_{i=1}^{N} x_i = \alpha \overline{x} \tag{5}$$

3. De même

$$\overline{x + \beta} = \frac{1}{N} \sum_{i=1}^{N} (x_i + \beta) = \frac{1}{N} \left(\sum_{i=1}^{N} x_i + \sum_{i=1}^{N} \beta \right) = \frac{1}{N} \sum_{i=1}^{N} x_i + \frac{1}{N} \sum_{i=1}^{N} \beta = \overline{x} + \beta \tag{6}$$

□

Les propriétés sont en fait bien connues : par exemple, la moyenne d'un ensemble de notes entre 0 et 20 est bien homogène à une note sur 20.

Remarque 1. Dans le cas où les données sont regroupées par valeurs (comme en (2)), la moyenne se définit aussi comme

$$\overline{x} = \frac{1}{N} \sum_{j=1}^{q} n_j v_j = \sum_{j=1} f_j v_j \tag{7}$$

ce qui revient à dire qu'on somme la valeur v_j le nombre n_j de fois où elle apparait ; quitte à faire passer le $1/N$ dans la somme, on retrouve alors la fréquence $f_j = n_j/N$.

III.2 La dispersion

Les indicateurs suivants étudient la **dispersion** de la série statistiques. Deux séries peuvent très bien avoir la même moyenne, et les mêmes minima et maxima, mais l'une où les valeurs sont très resserrées autour de la moyenne et l'autre où au contraire il y a soit des petites valeurs soit des grandes.

Définition 6. Soit une série statistique $(x_i)_{i=1,\dots,N}$ à une variable.

1. La **variance** de la série est la moyenne des carrés des écarts à la moyenne :

$$\boxed{s_x^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})^2} \tag{8}$$

C'est toujours nombre réel positif.

2. L'**écart-type** de la série est la racine carrée de la variance :

$$\boxed{s_x = \sqrt{s_x^2} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})^2}} \tag{9}$$

Propriétés 7. 1. La variance (de même pour l'écart-type) est nulle si et seulement si la série statistique est constante (ne prend qu'une seule valeur).

2. Si on multiplie toutes les valeurs de la série par une même constante $\alpha \in \mathbb{R}$, alors la variance est multipliée par α^2, et l'écart-type est multiplié par $|\alpha|$.

3. Si on additionne à toutes les valeurs de la série une même constante $\beta \in \mathbb{R}$, alors la variance et l'écart-type ne changent pas.

Démonstration. 1. Si la série $(x_i)_{i=1,\dots,N}$ est constante égale à une valeur x, alors sa moyenne $\overline{x}$ est aussi égale à x. Dans la somme tous les termes sont nuls, et donc $s_x^2 = 0$. Avec la racine carrée, s_x est nul si et seulement si s_x^2 est nul. Réciproquement, si la somme définissant s_x^2 est nulle alors comme c'est une somme de carrés (tous positifs) tous les termes sont nuls, donc pour tout i, $x_i = \overline{x}$: la série est constante, et donc égale nécessairement à sa moyenne.

2. Dans ce cas alors on sait déjà $\overline{\alpha x} = \alpha \overline{x}$ et donc

$$s_{\alpha x}^2 = \frac{1}{N}\sum_{i=1}^{N}(\alpha x_i - \alpha\overline{x})^2 = \frac{1}{N}\sum_{i=1}^{N}\alpha^2(x_i-\overline{x})^2 = \frac{1}{N}\alpha^2\sum_{i=1}^{N}(x_i-\overline{x})^2 = \alpha^2 \times \frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2 \tag{10}$$

On trouve donc $s_{\alpha x} = \sqrt{\alpha^2 s_x^2} = |\alpha|\sqrt{s_x^2} = |\alpha| s_x$.

3. Dans ce cas on sait que $\overline{x+\beta} = \overline{x} + \beta$ et alors

$$s_{x+\beta}^2 = \frac{1}{N}\sum_{i=1}^{N}\Big((x_i-\beta) - (\overline{x}+\beta)\Big)^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2 = s_x^2 \tag{11}$$

□

Remarque 2. Les deux premières propriétés ci-dessus sont une bonne motivation pour la définition de l'écart-type en sommant d'abord des écarts au carré, puis en prenant la racine carrée. En particulier par la deuxième propriété l'écart-type est *homogène*, s'exprime dans la même unité que les données de départ (l'écart-type de mesures en mètre s'exprime aussi en mètre, etc). Souvent on compte en terme de nombre d'écarts-types : les valeurs sont-elles écartés de la moyenne de une fois s_x, ou de deux fois s_x, ou de combien de fois s_x ? Dans beaucoup de situations naturelles (loi normales), elles sont rarement plus éloignées de trois fois l'écart-type.

Exercice 2. Une application bien utile de ces propriétés mathématiques est : supposons qu'après avoir corrigé un DS, on veuille fixer la moyenne à 10 et l'écart-type à 3. Alors on peut toujours y arriver en combinant la multiplication des notes par un coefficient (le même pour tous) et en ajoutant un certain nombre de points, à tous. Comment faire précisément ?

Remarque 3. On rencontre aussi une variance avec une division par $N-1$ et non pas N devant la somme. Ne pas poser de questions dessus à votre professeur.

Théorème 8 (Formule de König-Huygens)**.** *La variance se calcule plus facilement avec :*

$$\boxed{s_x^2 = \overline{x^2} - \overline{x}^2} \tag{12}$$

(la moyenne des carrés, moins le carré de la moyenne).

Démonstration. On écrit $s_x^2 = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}(x_i-\overline{x})^2$. Sous la somme, notons bien que la valeur $\overline{x}$ est une constante, on ne cherche surtout pas à la remplacer par sa définition. On développe alors le carré et on sépare les sommes :

$$\begin{aligned}
s_x^2 &= \frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2 && (13)\\
&= \frac{1}{N}\sum_{i=1}^{N}\left(x_i^2 - 2x_i\overline{x} + \overline{x}^2\right) && (14)\\
&= \frac{1}{N}\left(\sum_{i=1}^{N}x_i^2 - 2\overline{x}\sum_{i=1}^{N}x_i + \sum_{i=1}^{N}\overline{x}^2\right) && (15)\\
&= \underbrace{\frac{1}{N}\sum_{i=1}^{N}x_i^2}_{\overline{x^2}} - 2\overline{x} \times \underbrace{\frac{1}{N}\sum_{i=1}^{N}x_i}_{\overline{x}} + \underbrace{\frac{1}{N}\sum_{i=1}^{N}\overline{x}^2}_{\frac{1}{N}\times N\overline{x}^2} && (16)\\
&= \overline{x^2} - 2\overline{x}^2 + \overline{x}^2 && (17)\\
&= \overline{x^2} - \overline{x}^2
\end{aligned}$$

□

Pour s'en souvenir, il suffit de se souvenir du tout début de la preuve, c'est d'abord la somme des x_i^2 qui apparait. Si on se trompe de sens, on obtient de toute façon un résultat négatif.

Remarque 4. C'est cette formule qui est la plus pratique pour calculer la variance à la main : on écrit les valeurs de la série comme une liste et, en dessous, on écrit les carrés des valeurs. Puis on calcule chacune des deux moyennes et on applique la formule.

III.3 La position

Les indicateurs suivants sont bien connus aussi. L'idée tourne toujours autour de : mesurer la position et la dispersion de la série, de façon à ne pas être trop influencé par les valeurs extrêmes (trop grandes ou trop petites). Par exemple elles peuvent provenir d'erreurs de mesures, donc ne sont pas du tout à prendre un compte. Un autre exemple simple : dans une entreprise, il se peut que quelques dirigeants gagnent des millions d'euros, faisant augmenter le salaire moyen, mais cela ne dit rien du salaire des employés, qui préféreront se demander combien de personnes gagnent plus ou gagnent moins qu'eux… Les indicateurs ci-dessous sont tout à fait indépendants d'éventuelles valeurs extrêmes.

Définition 9. La **médiane** de la série statistique est *une* valeur Q_2 telle que la moitié de l'effectif a ses valeurs inférieures ou égales à Q_2.

La définition précise pose quelques petits problème. Si la série a un nombre impair de valeurs, alors on les range par ordre croissant et il y a exactement une valeur au milieu, c'est la médiane. Mais sinon, il y a deux valeurs au milieu x_k et x_{k+1} ($N = 2k$)

$$\underbrace{x_1 \leqslant x_2 \leqslant \cdots \leqslant x_k}_{N\text{valeurs}} \leqslant \underbrace{x_{k+1} \leqslant \cdots \leqslant x_{N-1} \leqslant x_N}_{N\text{valeurs}} \tag{18}$$

On peut alors :

1. Choisir le milieu $Q_2 = \frac{x_k + x_{k+1}}{2}$.
2. Ou bien choisir x_k. Dans ce cas, la médiane est la *plus petite valeur* Q_2 telle que *au moins* la moitié de l'effectif ait ses valeurs inférieurs ou égales à Q_2 (ce dernier « au moins » est nécessaire pour traiter les cas où les valeurs autour de la médiane sont toutes égales entre elles ; il n'est pas nécessaire si toutes les valeurs de la série sont différentes).

Cette deuxième définition se révèlera plus pratique à implémenter en informatique.
Si on a bien compris la médiane, alors la définition des quartiles est similaire :

Définition 10. On définit le **premier quartile** (resp. **troisième quartile**) comme une valeur Q_1 (resp. Q_3) tel que $\frac{1}{4}$ (resp. $\frac{3}{4}$) de l'effectif ait ses valeurs inférieures ou égales à Q_1 (resp. Q_3).

Là encore, si on trie la liste, cela ne tombe pas forcément sur une valeur exacte de la série. Alors :

1. Ou bien Q_1 devrait être compris entre deux valeurs consécutives x_r et x_{r+1}. Une bonne idée est alors de choisir non pas la moyenne de x_r et x_{r+1}, mais pondérée avec des coefficients $\frac{1}{2}$ ou $\frac{1}{4}$ ou $\frac{3}{4}$. C'est plus cohérent si les données sont régulièrement espacées, par exemple dans le cas extrême où on aurait seulement deux valeurs x_1, x_2, on voudrait que Q_1 soit proche à 1/4 de x_1 et à 3/4 de x_2 (donc c'est x_1 qui porte la pondération 3/4). Voir aussi la section IV.2.
2. Ou bien on définit Q_1 comme la plus petite valeur pour laquelle au moins un quart de l'effectif a ses valeurs inférieures à Q_1, et de même pour Q_3 (au moins trois quarts de l'effectif inférieur ou égal à Q_3), et c'est plus simple à écrire en informatique.

Plus généralement on peut prendre n'importe quelle proportion $p \in [0, 1]$, souvent exprimée en pourcentage, et considérer des **quantiles d'ordre** p. Cela intervient par exemple dans la phrase « Début 2021, 10 % des ménages français ont un patrimoine brut supérieur à 716 300 euros » (source : INSEE), on reconnait un quantile d'ordre 10 % (décile), en fait d'ordre 90 % si on regarde le patrimoine par ordre croissant.

Définition 11. L'**écart inter-quartiles** est la différence $Q_3 - Q_1$.

C'est un indicateur très général et très grossier de la dispersion de la série, comme l'écart-type, mais qui est complètement indépendant d'éventuelles valeurs extrêmes.

IV Implémentation en Python

IV.1 Les bases

Tout d'abord, tous les indicateurs que nous avons vus sont disponibles facilement avec Pandas. On les obtient d'un coup sur toute une colonne avec sa méthode `describe()` :

```
>>> tips["total_bill"].describe()
```

On peut les obtenir pour toutes les colonnes en même temps avec `tips.describe()`. On peut aussi obtenir ces indicateurs uns par uns, comme une méthode de l'objet `tips` et avec leur nom en anglais, par exemple `tips["total_bill"].mean()` :

- `count()` : effectif,
- `mean()` : moyenne,
- `min()`, `max()` : comme leur nom l'indique,
- `median()` : médiane,
- `std()` : écart-type, en anglais *standard deviation*.

Il n'y a semble-t-il pas de fonction directe pour les quartiles Q_1 et Q_3, mais une fonction `quantile(p)` prenant en argument une proportion p ($p = 0{,}25$ pour Q_1 et $p = 0{,}75$ pour Q_3).
Cependant, notre but est d'apprendre à les programmer à la main, sur des listes. Pour la première série, ce sont des très grands classiques.

Exercice 3. Écrire en Python les fonctions suivantes, prenant en argument une liste `L` représentant une série statistique à une variable (supposée non vide) :

1. `maximum(L)`, `minimum(L)`,
2. `moyenne(L)`,
3. `variance(L)`, `ecarttype(L)`. Pour utiliser la méthode de König-Huygens (et sourtout éviter de faire plusieurs boucles et de calculer plusieurs fois des moyennes !) on utilise une seule boucle, dans laquelle on calcule en même temps la somme des valeurs et la somme des carrés, puis on applique la formule de König-Huygens. La fonction racine carrée se nomme `sqrt` dans le module `numpy`.

La bibliothèque Pandas permet aussi d'obtenir ces indicateurs en *groupant* les données selon une certaine valeur d'un caractère qualitatif (analogue du `GROUP BY` en SQL), en appliquant les méthodes précédentes non pas à `tips` mais à `tips.groupby(`*`caractère`*`)`, par exemple :

```
>>> tips.groupby("sex")["total_bill"].describe()
        count       mean       std   min    25%    50%    75%    max
sex
Female   87.0  18.056897  8.009209  3.07  12.75  16.40  21.52  44.30
Male    157.0  20.744076  9.246469  7.25  14.00  18.35  24.71  50.81
```

Mathématiquement, cela signifie simplement qu'on a simplement séparé une série statistique en deux nouvelles séries, en fonction de la valeur de leur caractère `sex`.
Cela permet de répondre à des questions telles que :

Exercice 4. Qui semble en moyenne dépenser le plus, les hommes ou les femmes ? Quel jour de la semaine les dépenses sont les plus importantes ? Sont-elles en moyenne plus importants au déjeuner ou au dîner ? En facture totale, et en terme de pourboire ?

On veut maintenant s'intéresser au calcul de la médiane. C'est très facile quand les données sont déjà triées : la médiane est alors tout simplement l'élément au milieu, celui d'indice $n/2$ (à quelques détails près selon la parité de n) ! En général, cela ne marche pas si bien. On peut trier toute la liste, mais c'est long et compliqué. La méthode que nous proposons fonctionne seulement pour des valeurs **entières**, ainsi nous allons faire un détour en regroupant d'abord par classes entières.

IV.2 Regroupement en classes

Partant d'une série statistique à une variable $(x_i)_{i=1,\dots,N}$ présentant un caractère quantitatif, on peut vouloir définir des intervalles **disjoints** $[a_1, a_2[$, $[a_2, a_3[$, …, $[a_q, a_{q+1}[$ contenant toutes les valeurs de la série et **regrouper les données par classes** en indiquant pour chaque intervalle combien de valeurs sont dedans. Les données sont présentées comme un tableau :

valeur	$[a_1, a_2[$	$[a_2, a_3[$	$\cdots$	$[a_q, a_{q+1}[$
effectif	n_1	n_2	$\cdots$	n_q

$$\sum_{j=1}^{q} n_j = N \tag{19}$$

Après regroupement, la série se comporte un peu comme une série qualitative…

Remarque 5. Il est important que les intervalles recouvrent bien l'ensemble de toutes les valeurs de la série et soient disjoints (c'est donc une bonne idée de les prendre fermés à gauche et ouverts à droite comme ici ; ou l'inverse). Par exemple, dans un sondage contenant l'âge des individus, qu'on veut regrouper en disons 13–18 ans, 18–25 ans et 25–40 ans, il est important d'avoir bien précisé à l'avance dans laquelle ce ces catégories on range ceux qui ont exactement 18 ans ou bien 25 ans…

Réciproquement, étant donné un tel tableau, on peut effectuer divers calculs quantitatifs (moyenne, écart-type, etc) en faisant « comme si » les individus dans la classe $[a_j, a_{j+1}[$ avaient tous la même valeur qu'on choisit être le milieu $\frac{a_j+a_{j+1}}{2}$: c'est l'**approximation par le centre de la classe**. Si les données sont très nombreuses, on peut même faire l'approximation que nos données sont régulièrement espacées dans la classe, c'est-à-dire que les valeurs sont alignées selon une fonction affine ; la moyenne coïncide alors avec la médiane et avec le milieu de $[a_j, a_{j+1}[$. Appliquons ceci pour notre problème de factures et de pourboire. Pour la facture totale, nos classes seront simplement des nombres entiers. La fonction `int(x)` arrondit le réel x à l'entier « vers 0 » (si x est positif c'est bien la partie entière). Il faudra passer à notre fonction une liste `L` mais aussi une borne maximale sur les valeurs de la série qu'on étudie.

Exercice 5. Écrire une fonction `classes_entieres(L, b)` qui prend en argument une liste `L` (on suppose que ce sont des valeurs réelles positives) et un entier b (on suppose que toutes les valeurs de `L` sont inférieures ou égales à b) et qui renvoie une liste `C` de longueur b telle que `C[j]` compte le nombre de valeurs de `L` qui sont dans l'intervalle $[j, j+1[$.

On pourrait ajuster la fonction pour des classes plus petites (par exemple sautant de 0,5) ou au contraire plus larges.

IV.3 Algorithme de calcul de la médiane

La médiane, ainsi que les quartiles, se lisent rapidement quand on part d'une série de nombres entiers comme la liste `C` ci-dessus et qu'on forme la liste des **effectifs cumulés croissants** : une liste `F` de même longueur que `C` dont le j-ème élément donne le nombre de valeurs de la série qui sont inférieurs ou égales à j. On les calcule de proches en proches. Il suffira alors de chercher à partir de quel indice de `F` on dépasse la moitié de l'effectif ! Le dernier élément de `F` est l'effectif total, c'est-à-dire la somme de toutes les valeurs de `C`.

Exercice 6. Écrire une fonction `cumul(C)` qui prend en argument une liste `C`, supposée de nombres entiers positifs, et qui renvoie une liste `F` de même longueur telle que `F[j] = C[0] + ... + C[j]`. Les valeurs de `F` se calculent de proche en proche, à partir de `F[0] = C[0]`.

On peut alors lire la médiane d'une liste `L`, en prenant son regroupement en classes `C` comme ci-dessus puis en construisant la liste `F`. On cherche alors avec une boucle le plus petit indice de `F` où la valeur dépasse la moitié de l'effectif total.

Exercice 7. Écrire cette fonction `mediane(C)`, dont l'argument est `C`.

Après coup, on voit qu'on peut se passer de construire la liste `F`, mais calculer les effectifs cumulés au fur et à mesure dans la boucle.

Exercice 8. Écrire la fonction `quantile(C, p)`, puis en déduire les fonctions `quartile1(C)`, `quartile3(C)` et `ecart_inter_quartiles(C)`.

V Représentation graphique

On s'intéresse à divers modes de représentation graphique des données. Cela dépend fortement du choix des données qualitatives ou quantitatives.
Les diverses représentations ci-dessous sont à tester dans le fichier matériel. Les fonctions existent sous la forme de méthodes des objets Pandas, mais tout le tracé est en réalité passé à la bibliothèque Matplotlib, d'ailleurs, les fonctions existent dans Matplotlib en prenant en argument une ou deux listes (ou tableau Numpy) de valeurs. Et très certainement aucune connaissance spécifique à ces bibliothèques n'est à exiger.

Diagramme en barres Permet de visualiser très simplement des données **qualitatives**, en abscisse les valeurs uniques de la série (les « étiquettes » des valeurs qualitatives), et verticalement une barre proportionnelle à l'effectif ou bien à la fréquence.

Diagramme en camembert En anglais c'est un *pie chart* (tarte). Similairement au diagramme en barres, on part de données qualititaves regroupées par valeurs uniques, et le secteur découpé sur le cercle est proportionnel à l'effectif. Dans ce cas l'effectif total correspond à un angle de 2π soit 360 degrés ; écrivant une banale proportionnalité, on calcule les angles de chaque secteur.

Boite à moustaches C'est une représentation d'une **série statistique quantitative à une variable**. Les valeurs (nombres réels) sont représentées alignées, une boite entoure les valeurs entre les quartiles Q_1 et Q_3 avec un trait au milieu représentant la médiane Q_2. Les moustaches de la boite s'étendent, diverses conventions existent (par exemple : s'étendre du premier au neuvième décile, laissant donc 10 % des valeurs en dehors d'un côté et de même 10 % de l'autre), les valeurs extrêmes sont éventuellement indiquées individuellement.

Histogramme Ressemble a priori au diagramme en barres, mais pour des données **quantitatives** qui sont **regroupées par classes** sur l'axe des abscisses. En ordonnée, l'effectif de la classe. Les rectangles sont donc naturellement collés les uns aux autres, la finesse des classes est choisie pour que l'ensemble des rectangles dessine une courbe à peu près lisse : plus elles sont fines, plus la courbe va se lisser, mais si elles sont trop fines, elles risquent de contenir trop peu de valeurs et donc on voit apparaitre trop d'anomalies.

Courbe des effectifs cumulés Pour une série quantitive x, il s'agit du graphe de la fonction réelle $F_x : t \mapsto$ nombre de valeurs qui sont inférieures ou égales à t. La fonction F_x a les propriétés suivantes :

1. F_x est croissante,
2. $\lim_{t\to-\infty} F_x(t) = 0$ et $\lim_{t\to+\infty} F_x(t) = N$ (effectif total),
3. F est une fonction constante par morceaux et continue à droite (comme la fonction partie entière), en effet elle change brusquement de valeur précisément quand t passe sur une des valeurs de la série x,
4. Mais on peut aussi remplacer les segments de courbes horizontaux par des droites de telle façon à obtenir une fonction *continue* (ligne polygonale) ; dans le cas notamment de données regroupées en classe, cela revient à supposer que les données sont uniformément réparties dans la classe, et donc $F_x(t)$ croît dans la classe comme une fonction affine,
5. La médiane se lit alors comme la valeur de t telle que $F_x(t) = N/2$.

Il existe aussi une courbe des fréquences cumulées, de la même manière mais avec les fréquences ; dans ce cas $\lim_{t\to+\infty} F_x(t) = 1$, et la médiane est la valeur de t telle que $F_x(t) = 1/2$.

Nuage de points Voir la partie VI, cela concerne réellement les séries statistiques à deux variables.

VI Statistiques à deux variables

On suppose maintenant qu'on a une série statistique à deux variables. Dans une population de taille N, on mesure donc simultanément deux caractères quantitatifs x et y, et on note x_i, y_i les valeurs pour l'individu i.
On peut alors s'intéresser à tous les indicateurs précédents pour chacune des deux séries séparément.

Définition 12. Le **nuage de points** de la série à deux variable est l'ensemble des points de coordonnée (x_i, y_i), pour $i = 1, \dots, N$.

Comme son nom l'indique, il s'agit d'un ensemble de points du plan, sans aucune nécessité que cela ne trace une figure particulière.

Définition 13. Le point de coordonnées $(\overline{x}, \overline{y})$ s'appelle **point moyen**.

On peut montrer que c'est le barycentre en physique...
On obtient le nuage de points avec Pandas et le type de graphique `scatter`, en lui donnant le nom des deux séries :

```
>>> tips.plot.scatter("total_bill", "tip")
>>> plt.show()
```

On se doute bien qu'il y a un lien entre ces deux séries : plus la facture est grande, plus le pourboire sera grand. Un nouvel indicateur fait son entrée en jeu.

Définition 14. La **covariance** de la série statistique à deux variables (x, y), d'effectif total N, est le nombre

$$\boxed{s_{x,y} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})} \tag{20}$$

Si $x = y$ alors on retrouve la définition de la variance s_x^2. En général nous allons voir que la covariance est une mesure de la dépendance entre deux séries statistiques. Observez aussi que la définition est symétrique en x et y : $s_{x,y} = s_{y,x}$. Observez la notation : les quantités avec soit x, y en bas, soit x en bas et un carré en haut, sont bien homogènes à des produits ou des carrés ; alors que s_x s'obtient avec une racine carrée sur $s_x^2 = s_{x,x}$.

Exercice 9. Écrire en Python la fonction `covariance(X, Y)`, prenant en argument deux listes supposées de même longueur représentant une série statistique à deux variables.

Proposition 15. *Si y est lié à x par une relation affine, c'est-à-dire s'il existe $(\alpha, \beta) \in \mathbb{R}^2$ tels que $\forall 1 \leqslant i \leqslant N$, $y_i = \alpha x_i + \beta$ (ce qu'on note $y = \alpha x + \beta$), alors* $\boxed{s_{x,y} = \alpha s_{x,x} = \alpha s_x^2}$.

Démonstration. Dans ce cas $\overline{y} = \alpha\overline{x} + \beta$ et donc

$$\begin{aligned}
s_{x,y} &= \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})((\alpha x_i + \beta) - (\alpha\overline{y} + \beta)) && (21)\\
&= \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(\alpha x_i - \alpha\overline{y}) && (22)\\
&= \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x}) \times \alpha(x_i - \overline{y}) && (23)\\
&= \frac{1}{N}\alpha\sum_{i=1}^{N}(x_i - \overline{x}) \times (x_i - \overline{y}) && (24)\\
&= \alpha s_{x,y} && \square
\end{aligned}$$

Théorème 16 (König-Huygens pour la covariance)**.** $s_{x,y} = \overline{x \times y} - \overline{x} \times \overline{y}$

Démonstration. Exercice, même méthode que pour la variance (théorème 8). □

Une formule intéressante motivant l'utilisation de carrés dans la définition de la variance : pouvoir s'intéresser à la variance de somme de deux séries statistiques (la série des $x_i + y_i$).

Proposition 17. *On a le développement suivant : $s_{x+y}^2 = s_x^2 + 2s_{x,y} + s_y^2$. Ainsi, si $s_{x,y} = 0$ alors $s_{x+y}^2 = s_x^2 + s_y^2$.*

Démonstration. Une démonstration purement avec König-Huygens : on a $\overline{x+y} = \overline{x} + \overline{y}$, et donc

$$\begin{aligned} s_{x+y}^2 &= \overline{(x+y)^2} - \overline{x+y}^2 && (25)\\ &= \overline{x^2+2xy+y^2} - \left(\overline{x}+\overline{y}\right)^2 && (26)\\ &= \left(\overline{x^2}+2\overline{xy}+\overline{y^2}\right) - \left(\overline{x}^2+2\overline{x}\times\overline{y}+\overline{y}^2\right) && (27)\\ &= \overline{x^2}+2\overline{xy}+\overline{y^2}-\overline{x}^2-2\overline{x}\times\overline{y}-\overline{y}^2\Big) && (28)\\ &= \left(\overline{x^2}-\overline{x}^2\right)+2\left(\overline{xy}-\overline{x}\times\overline{y}\right)+\left(\overline{y^2}-\overline{y}^2\right) && (29)\\ &= s_x^2+2s_{x,y}+s_y^2 \end{aligned}$$ □

On peut aussi écrire la démonstration avec des sommes. Cette formule ressemble à une formule de géométrie sur la norme des vecteurs $\|\overrightarrow{u+v}\|^2 = \|\overrightarrow{u}\|^2 + 2\overrightarrow{u}\cdot\overrightarrow{v} + \|\overrightarrow{v}\|^2$, illustre l'intérêt de ces méthodes de sommer des carrés ; $s_{x,y}$ se comporte un peu comme un produit scalaire entre deux séries, et s_x^2 comme une norme au carré…

Définition 18. Soit une série statistique à deux variables (x, y). On suppose que les deux séries x et y ne sont pas constantes. Le **coefficient de corrélation linéaire** de la série statistique à deux variables (x, y) est le nombre

$$\boxed{r_{x,y} = \frac{s_{x,y}}{s_x s_y}} \tag{30}$$

Ce nombre ne dépend pas de l'ordre de x et y.

Remarque 6. Si y est lié à x par la relation affine $y = \alpha x + \beta$, alors à cause de la proposition 15, on a $s_{x,y} = \alpha s_x^2$ et $s_y = |\alpha| s_x$, donc $r_{x,y} = \frac{\alpha}{|\alpha|} = \begin{cases} 1 & \text{si } \alpha > 0 \\ -1 & \text{si } \alpha < 0 \end{cases}$.

Théorème 19 (Inégalité de Cauchy-Schwarz)**.** *Le coefficient vérifie $r_{x,y} \in [-1, 1]$. De plus il est égal à -1 ou $+1$ si et seulement si il existe une relation affine $y = \alpha x + \beta$ entre les deux séries ; plus précisément il est de 1 si $\alpha > 0$ (séries positivement corrélées : plus x augmente, plus y augmente aussi) et -1 si $\alpha < 0$ (négativement corrélées : plus x augmente, plus y diminue).*

Démonstration. On souhaite démontrer $\left|\frac{s_{x,y}}{s_x s_y}\right| \leqslant 1$, c'est-à-dire $|s_{x,y}| \leqslant |s_x| \times |s_y|$. Comme tout est positif, cela est aussi équivalent à $(s_{x,y})^2 \leqslant s_x^2 s_y^2$. La démonstration est analogue à celle de l'inégalité démontrée sur les vecteurs $|\overrightarrow{u}\cdot\overrightarrow{v}|^2 \leqslant \|\overrightarrow{u}\|^2 \times \|\overrightarrow{v}\|^2$.
Dans le cas $s_x = 0$ alors la série statistique x est constante, et on vérifie tout de suite $s_{x,y} = 0$ aussi, il n'y a rien à faire. De même dans le cas $s_y = 0$.
Sinon, on considère la fonction

$$P : t \in \mathbb{R} \longmapsto s_{tx+y}^2 \tag{31}$$

Alors P se développe par la proposition 17 en

$$\begin{aligned} \forall t \in \mathbb{R}, \quad P(t) &= s_{tx}^2 + 2s_{tx,y} + s_y^2 && (32)\\ &= t^2 s_x^2 + 2ts_{x,y} + s_y^2 && (33) \end{aligned}$$

On reconnait une fonction polynôme du degré 2 (seul t varie, tous les autres termes sont des constantes), de la forme $at^2 + bt + c$ avec $a = s_x^2$, $b = 2s_{x,y}$ et $c = s_y^2$. De plus, elle ne prend que des valeurs positives, puisque par définition c'est une variance. Donc son discriminant $\Delta = a^2 - 4ac$ est négatif (pas de changement de signe de la fonction), ce qui donne

$$4s_{x,y}^2 - 4s_x^2 s_y^2 \leqslant 0 \tag{34}$$

c'est-à-dire $s_{x,y} \leqslant s_x^2 s_y^2$.
De plus, il y a égalité si et seulement si $\Delta = 0$. Dans ce cas le polynôme P admet une unique racine (double), donc il existe $t_0 \in \mathbb{R}$ tel que $s_{t_0 x+y}^2 = 0$. Cela signifie que la série des $t_0 x + y$ est constante. Posant β la valeur de cette constante (qui est aussi, du coup, la moyenne de la série), alors $t_0 x + y = \beta$ donc $y = -t_0 x + \beta$, et ceci est notre relation affine (quitte à poser $\alpha = -t_0$). □

Définition 20. Soit une série statistique à deux variables (x, y), on suppose $s_x^2 \neq 0$ (sinon, les valeurs selon x sont constantes et le nuage de points est aligné verticalement). La **droite de régression affine** de la série statistiques à deux variables (x, y) est la droite, dont on note en majuscule les variables,

$$Y = \frac{s_{x,y}}{s_x^2}(X - \overline{x}) + \overline{y} \tag{35}$$

C'est la droite qui passe par le point moyen $(\overline{x}, \overline{y})$ et de coefficient directeur $\frac{s_{x,y}}{s_x^2}$.

On pourra démontrer que cette droite minimise la quantité $\sum_{i=1}^{N} \Big((\alpha x_i + \beta) - y_i\Big)^2$ pour $(\alpha, \beta) \in \mathbb{R}^2$: parmi toutes les droites d'équation $Y = \alpha X + \beta$, c'est celle qui est en un sens la « plus proche » du nuage de points, où la proximité se définit en sommant les écarts *verticaux*, au carré, entre la droite qui passe par (x_i, Y) et le point (x_i, y_i).

Remarque 7. Dans le cas où la droite passe exactement par ces points, alors on obtient une relation affine entre x et y, notée $y = \alpha x + \beta$. Mais si une telle relation existe, alors $s_{x,y} = \alpha s_x^2$, et donc le coefficient α est nécessairement égal à $\frac{s_{y,y}}{s_x^2}$. De plus, dans ce cas, la droite passe nécessairement par le point moyen (c'est une propriété de base des fonctions affines, la généralisation de : si elle passe par deux points, alors elle passe par le milieu). On peut donc penser à la droite de régression affine comme à la droite exprimée en terme de covariance qu'on obtiendrait si les séries étaient en relation affine.

Exercice 10.

1. Donner les coefficients de la droite de régression affine entre le pourboire et la facture.
2. Interprétation ?
3. Représenter sur un graphique simultanément le nuage de points et la droite de régression (tracée avec Matplotlib comme d'habitude).