TP 3

Statistiques descriptives univariées

Compétences attendues.

- ✓ Regrouper une série statistique par modalités ou par classes.
- ✓ Connaître les indicateurs de position (moyenne, médiane, quartiles) et les commandes associées.
- ✓ Connaître les indicateurs de dispersion (écart-type, étendue, distance inter-quartile) et les commandes associées.
- ✓ Représenter graphiquement une série statistique.

Liste des commandes `Python` exigibles aux concours.

- Dans la librairie `numpy` : `np.sum, np.min, np.max, np.cumsum, np.mean, np.median, np.var, np.std`.
- Dans la librairie `matplotlib.pyplot` : `plt.hist, plt.bar, plt.boxplot`.
- Dans la librairie `pandas` : `pd.read_csv, pd.head, pd.shape, pd.discribe, pd.mean, pd.median, pd.var, pd.std, pd.count, pd.sort_values`

Anthony Mansuy
Professeur de Mathématiques en deuxième année de CPGE filière ECG au Lycée Clemenceau (Reims)
Page personnelle : `http://anthony-mansuy.fr`
E-mail : `mansuy.anthony@hotmail.fr`

L'objet des statistiques descriptives univariées (ou unidimensionnelles) est de fournir des résumés synthétiques, graphiques et numériques, de séries de valeurs observées sur une population ou un échantillon. On présente ici les indicateurs les plus couramment employés pour décrire une série statistique.

L'étude statistique ressemble beaucoup à la théorie des probabilités. La différence fondamentale entre statistiques et probabilités est la suivante :

- En probabilités, on cherche à anticiper l'avenir (en calculant nos chances de gagner à un jeu avant de jouer par exemple), c'est-à-dire à obtenir des informations avant d'effectuer une expérience aléatoire. On parle de calculs théoriques (qui induisent/anticipent les résultats).
- En statistiques, on cherche à obtenir des informations après avoir effectué des expériences, souvent après avoir répété plusieurs fois la même expérience. On parle de calculs empiriques (déduits de l'expérience).

1 Principales notions en statistiques descriptives

1.1 Présentation des données

On considère un ensemble Ω appelé **population** en statistique descriptive. On appellera ses éléments ω des **individus**.

Exemple. Ω = l'ensemble de la population française, Ω = l'ensemble des voitures immatriculées en France.

On étudie un **caractère** de cette population :

Définition.

Un **caractère** (ou **variable**) sur la population Ω est une application $X : \Omega \to E$, où E désigne un ensemble quelconque.

Si E est un ensemble de nombres, on dit que X est un caractère **quantitatif**. Dans le cas contraire, on parle de caractère **qualitatif**.

Exemple. Un caractère possible sur la population française est la taille (caractère quantitatif) ou encore la couleur des yeux (caractère qualitatif).

Nous ne traiterons que du cas des caractères quantitatifs.

Contrairement aux probabilités, nous allons observer les valeurs prises par la variable X sur une grande population (c'est-à-dire simuler un grand nombre de fois la variable X) et obtenir des informations sur X grâce à ces simulations (loi empirique : tableau des fréquences, moyenne empirique...) au lieu de l'étudier théoriquement avant de faire une expérience.

Pour obtenir un renseignement exact sur un caractère X, il faudrait étudier tous les individus de la population Ω. Lorsque cela n'est pas possible, on étudie seulement les individus d'une partie finie $\{\omega_1, \dots, \omega_n\}$ de Ω appelée **échantillon observé**. Son cardinal n est alors la **taille** ou l'**effectif** de l'échantillon.

Définition.

- On appelle **série statistique** d'un échantillon $\{\omega_1, \dots, \omega_n\}$ de Ω pour le caractère X la donnée de la liste $(x_1, \dots, x_n) = (X(\omega_1), \dots, X(\omega_n))$ des valeurs prises par X sur l'échantillon.
- Les valeurs m_i prises par X sont appelées **modalités**.
- L'**effectif d'une modalité** m_i est le nombre n_i de fois où m_i apparait dans la série statistique $(x_1, \dots, x_n)$.
- La **fréquence d'une modalité** m_i est le réel $f_i = \dfrac{\text{effectif}}{\text{effectif total}} = \dfrac{n_i}{n}$.
  En pratique, c'est le taux de la population dont le caractère X prend la valeur m_i.
- La **fréquence cumulée d'une modalité** m_i le réel $p_i = \displaystyle\sum_{m_j \leq m_i} f_j$.
  En pratique, c'est le taux de la population dont le caractère X prend une valeur inférieure ou égale à la modalité m_i.

Remarques.

1. Si $x = (x_1, \dots, x_n)$ est une série statistique, $(m_1, \dots, m_p)$ ses modalités, $(n_1, \dots, n_p)$ ses effectifs et $(f_1, \dots, f_p)$ ses fréquences, alors on a :

$$\sum_{i=1}^{p} n_i = n \qquad \text{et} \qquad \sum_{i=1}^{p} f_i = \sum_{i=1}^{p} \frac{n_i}{n} = \frac{\sum_{i=1}^{p} n_i}{n} = 1$$

2. Les notions suivantes se correspondent en probabilités et en statistiques :

X variable aléatoire	$\leftrightarrow$	X variable statistique
Support $X(\Omega)$	$\leftrightarrow$	L'ensemble des modalités m_i
Probabilité $P(X = x_i)$	$\leftrightarrow$	Fréquence f_i
Fonction de répartition F_X	$\leftrightarrow$	Fréquence cumulée p_i

Définition.

Pour présenter les données d'une série statistique, on peut effectuer :

- Un **regroupement par modalités** (dans le cas où le nombre de modalités est faible) :

  On regroupe la série statistique par modalités - effectifs, c'est-à-dire qu'on donne :

  - la liste (m_i) des modalités du caractère X,
  - les effectifs (n_i) correspondants.

  On peut aussi choisir de présenter cette série regroupée par modalité - fréquence, en donnant les modalités (m_i) et les fréquences des modalités (f_i) correspondantes.

- Un **regroupement par classes** (dans le cas où le nombre de modalités est grand) :

  Plutôt que de conserver toutes les valeurs, il est plus intéressant de les regrouper par classes :

  - on considère une suite de réels $c = (c_0 < \dots < c_k)$ définissant les **classes** $I_1 = [c_0, c_1]$, $I_2 =]c_1, c_2]$, $\dots$, $I_k =]c_{k-1}, c_k]$, l'**amplitude** de la classe I_i étant $c_i - c_{i-1}$;
  - On note n_i le nombre d'éléments de X appartenant à l'intervalle I_i pour $1 \leq i \leq k$.

  On se ramène ainsi à une série statistique de taille k, dont les modalités sont les milieux $y_i = \dfrac{c_{i-1} + c_i}{2}$ des classes et d'effectifs correspondants les n_i.

1.2 Indicateurs de position

Définition.

On appelle **moyenne empirique** de la série statistique $x = (x_1, \dots, x_n)$ le réel :

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Remarques.

1. Si la série statistique x est groupée par modalités - effectifs, avec les modalités $(m_1, \dots, m_p)$ d'effectifs $(n_1, \dots, n_p)$ et de fréquences $(f_1, \dots, f_p)$, alors on a :

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{p} m_i \cdot n_i = \sum_{i=1}^{p} m_i \cdot \frac{n_i}{n} = \sum_{i=1}^{p} m_i \cdot f_i.$$

2. Les notions d'espérance en probabilités et de moyenne en statistiques se correspondent :

$$E(X) = \sum_{x_i \in X(\Omega)} x_i P(X = x_i) \quad \leftrightarrow \quad \overline{x} = \sum_{i=1}^{p} m_i \cdot f_i.$$

Définition.

La **médiane** d'une série statistique ordonnée est un réel m partageant la série en deux séries d'effectifs égaux. Si $(x_1 \leq x_2 \leq \cdots \leq x_n)$ est la série statistique ordonnée, m est défini par :

- si $n = 2p - 1$ est impaire, $m = x_p$ (la valeur du milieu) ;
- si $n = 2p$ est paire, $m = \dfrac{x_p + x_{p+1}}{2}$ (la moyenne des deux termes du milieu).

Remarque. La médiane ne s'intéresse qu'à la valeur "centrale", sans tenir compte des valeurs extrémales. La moyenne est au contraire "déformée" par les valeurs extrémales.

Propriété 1 (Moyenne et médiane)

Soit `x` un vecteur.

- `np.mean(x)` donne la moyenne du vecteur `x`.
- `np.median(x)` donne une médiane du vecteur `x` (non nécessairement ordonné).

Définition.

Soit $x = (x_1, \ldots, x_n)$ une série statistique.

- Le **premier quartile** q_1 de x est la plus petite valeur de x telle que 25 % des valeurs lui soient inférieures ou égales.

  C'est donc la médiane de la sous-série statistique formée en ne gardant que la première moitié des valeurs x_i rangées dans l'ordre croissant.
- Le **troisième quartile** q_3 de x est la plus petite valeur de x telle que 75 % des valeurs lui soient inférieures ou égales.

  C'est donc la médiane de la sous-série statistique formée en ne gardant que la seconde moitié des valeurs x_i rangées dans l'ordre croissant.

Remarque. De même, on définit les **déciles** et les **centiles** d'une série statistique :

- Pour $k \in [\![1, 99]\!]$, le k-ième centile est la valeur c_k de la série pour laquelle moins de k % de la population prend des valeurs strictement inférieures à c_k et moins de $(100 - k)$ % de la population prend des valeurs strictement supérieures à c_k.
- Pour $k \in [\![1, 9]\!]$, le k-ième décile est la valeur d_k de la série pour laquelle moins de k % de la population prend des valeurs strictement inférieures à d_k et moins des $(10 - k)$ dixièmes de la population prend des valeurs strictement supérieures à d_k.

Définition.

- Si une série statistique est regroupée par modalité, on appelle alors **mode** toute modalité pour laquelle l'effectif est maximal (il peut y en avoir plusieurs).
- Si une série statistique est regroupée par classes, on appelle alors **classe modale** toute classe correspondant au rectangle de plus grande hauteur de l'histogramme de cette série.

1.3 Indicateurs de dispersion

Définition.

Soit $x = (x_1, \dots, x_n)$ une série statistique.

- On appelle **variance empirique** de x le nombre réel positif :

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

- On appelle **écart-type empirique** de x le réel s_x.

Remarques.

1. Comme en probabilités, la variance mesure la dispersion des valeurs de x par rapport à sa moyenne.
2. Comme en probabilités, la formule de Koenig-Huygens est valable :

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\overline{x} + \overline{x}^2) = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2 \times \overline{x} \times \underbrace{\frac{1}{n}\sum_{i=1}^{n}x_i}_{=\overline{x}} + \overline{x}^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2.$$

3. Si la série statistique x est groupée par modalités - effectifs, avec les modalités $(m_1, \dots, m_p)$ d'effectifs $(n_1, \dots, n_p)$ et de fréquences $(f_1, \dots, f_p)$, alors on a :

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{p}(m_i - \overline{x})^2 \cdot n_i = \sum_{i=1}^{p}(m_i - \overline{x})^2 \cdot \frac{n_i}{n} = \sum_{i=1}^{p}(m_i - \overline{x})^2 \cdot f_i.$$

4. Les notions de variance et d'écart type en probabilités et en statistique se correspondent :

$$V(X) = \sum_{x_i \in X(\Omega)}(x_i - E(X))^2 P(X = x_i) \quad \leftrightarrow \quad s_x^2 = \sum_{i=1}^{p}(m_i - \overline{x})^2 \cdot f_i$$
$$\sigma(X) = \sqrt{V(X)} \quad \leftrightarrow \quad s_x$$

Définition.

- On appelle **étendue** d'une série statistique la différence entre la plus grande et la plus petite modalité.
- On appelle **distance inter-quartile** le réel $q_3 - q_1$.

Remarque. La distance inter-quartile est un indicateur de dispersion : c'est la longueur de l'**intervalle inter-quartile** $[q_1, q_3]$, lequel contient la moitié des valeurs de la série, réparties autour de la médiane m.

Propriété 2 (Variance, écart-type et étendue)

Soit `x` un vecteur.

- `np.var(x)` donne la variance du vecteur `x`.
- `np.std(x)` (pour standard deviation) donne l'écart-type du vecteur `x`.
- `np.max(x)-np.min(x)` donne l'étendue du vecteur `x`.

Remarques.

1. On peut aussi obtenir la variance de `x` avec la commande `v = np.mean((x-np.mean(x))**2)` (en utilisant sa définition) ou indifféremment `v = np.mean(x**2)-np.mean(x)**2` (par Koenig-Huygens).
2. On peut aussi déduire l'écart-type de `x` à partir de sa variance `v` avec la commande `np.sqrt(v)` (en utilisant sa définition).

Propriété 3 (Transformation affine)

Soit $x = (x_1, \dots, x_n)$ une série statistique de moyenne $\overline{x}$, de médiane m, de variance v et d'écart type σ.
Considérons la série statistique obtenue en transformant chaque x_i en $ax_i + b$, où a et b sont deux réels. Alors, la moyenne, la médiane, la variance et l'écart type de cette nouvelle série statistique sont :

$$a \cdot \overline{x} + b, \quad a \cdot m + b, \quad a^2 \cdot v \quad \text{et} \quad |a| \cdot \sigma.$$

2 Représentations graphiques

2.1 Diagrammes en bâtons

Définition.

On représente une série statistique **groupée par modalités** en plaçant sur l'axe horizontal les modalités et en dressant à la verticale de chacune un bâton de hauteur égale à son effectif ou sa fréquence (resp. son effectif cumulé ou sa fréquence cumulée).

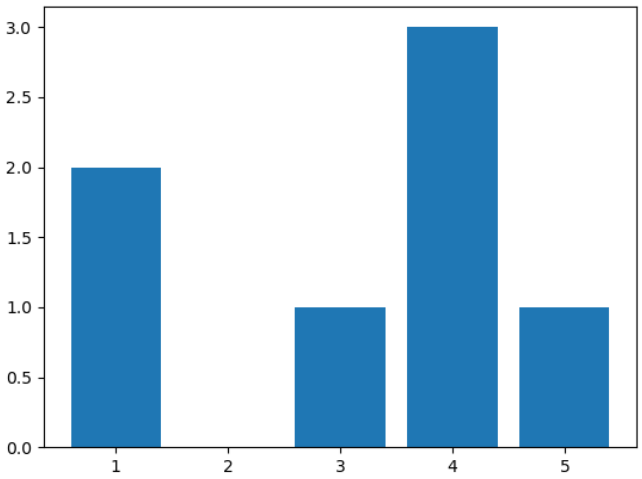
Propriété 4 (Diagrammes en bâtons)

Soit `m` et `n` deux vecteurs de même longueur.
L'instruction `plt.bar(m,n)` commande le tracé du diagramme en bâtons associé à la série statistique, la liste `m` définissant les modalités distinctes d'une série statistique (représentés en abscisse) et `n` les effectifs associés (représentés par la hauteur des bâtons).

Exemple. En entrant les instructions suivantes dans la console,

```
>>> m = np.array([1, 3, 4, 5])
>>> n = np.array([2, 1, 3, 1])
>>> plt.bar(m,n)
>>> plt.show()
```

on obtient le diagramme en bâtons :



2.2 Histogrammes

Définition.

On représente une série statistique **groupée par classes** en plaçant les c_i sur un axe horizontal et en traçant à la verticale un rectangle de base $[c_i, c_{i+1}]$ d'aire égale à la fréquence de la classe correspondante.

Propriété 5 (Histogrammes)

Soit `x` un vecteur.

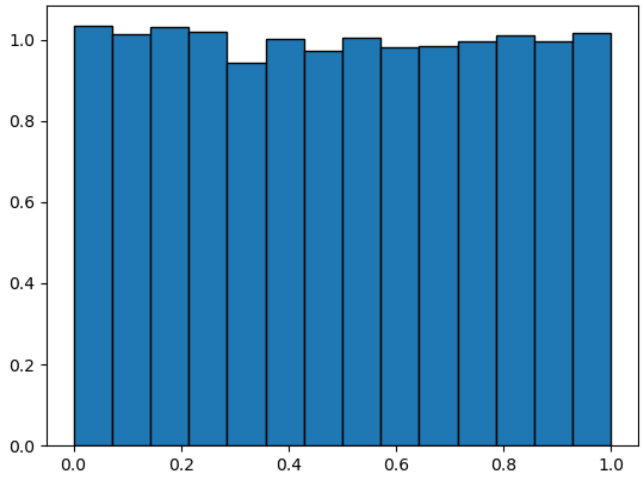
- L'instruction `plt.hist(x,n)` commande le tracé de l'histogramme associé à la série `x` en `n` classes découpées entre la plus petite valeur de `x` et la plus grande (par défaut, `n` vaut 10).
- Si `c` contient un vecteur $(c_1, \dots, c_m)$, l'instruction `plt.hist(x,c)` définit les classes à l'aide de `c` : la i-ième classe a pour extrémités c_i et c_{i+1}.

Remarque. `plt.bar` et `plt.hist` possèdent un grand nombre d'options dont aucune n'est exigible mais qui peuvent être pratique pour rendre les schémas lisibles (notamment `legend` qui s'emploie comme pour les courbes de fonctions, `density = 'True'` qui calibre les rectangles pour que le total de leurs surfaces soit égales à 1 et `edgecolor = 'k'` qui permet de délimiter les rectangles en noir).

Exemple. En entrant les instructions suivantes dans la console,

```
>>> x = rd.random(10000)
>>> cl = np.linspace(0,1,15)
>>> plt.hist(x, cl, density = 'True', edgecolor = 'k')
>>> plt.show()
```

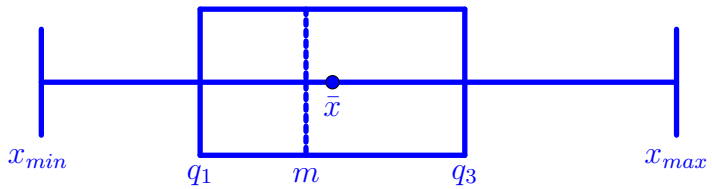
on obtient l'histogramme :



2.3 Boites à moustaches

Définition.

La **boîte à moustache** d'une série statistique est un schéma permettant de visualiser l'étendu de la série (la plus petite valeur et la plus grande), les valeurs des quartiles q_1 et q_3 entre lesquelles se concentrent la moitié de la masse ainsi que la médiane :



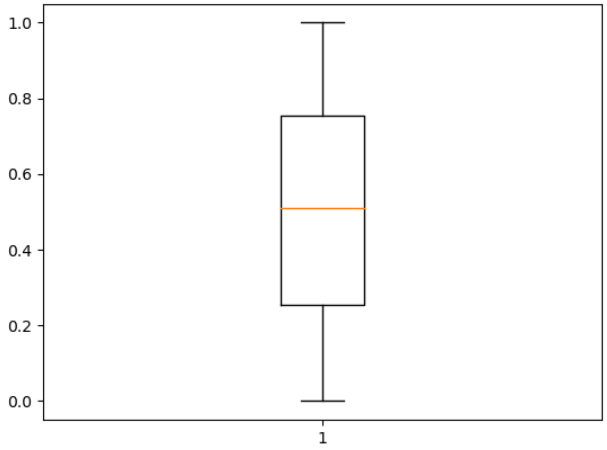
Propriété 6 (Boites à moustaches)

Soit `x` un vecteur.
L'instruction `plt.boxplot(x)` commande le tracé de la boîte à moustache associée à `x`.

Exemple. En entrant les instructions suivantes dans la console,

```
>>> x = rd.random(10000)
>>> plt.boxplot(x)
>>> plt.show()
```

on obtient la boite à moustaches :



Exercice 1 (★)

1. On souhaite étudier le nombre de buts par match de foot durant un tournoi.

   Préciser la population Ω étudiée et le caractère observé X.

2. Pour éviter d'avoir à regarder tous les matchs du tournoi, nous allons faire une étude sur un échantillon de 20 matchs. Taper dans la console l'instruction suivante donnant les nombres de buts marqués lors de ces matchs :

   ```
   >>> x = np.array([3, 1, 5, 3, 2, 7, 0, 1, 0, 3, 2, 4, 4, 0, 3, 3, 2, 5, 3, 1])
   ```

3. Avec `Python`, calculer la moyenne de buts par match, la médiane, la variance, l'écart type et l'étendue.

4. Compléter "à la main" le tableau suivant :

Modalités							
Effectifs							
Fréquences							
Fréquences cumulées							

5. Déduire du tableau précédent le mode, le premier, le deuxième et le troisième quartile.

6. Définir sur `Python` deux vecteurs `m` et `n` représentant les modalités et les effectifs. En déduire les vecteurs `f` et `fcc` des fréquences et des fréquences cumulées.

7. Entrer les instructions `np.sum(n)` et `np.sum(f)`. Quelles sont les valeurs renvoyées ? Pourquoi ?

8. Tracer les diagrammes en bâtons des effectifs, des fréquences et des fréquences cumulées.

9. Tracer la boîte à moustaches associée à `x`.

Exercice 2 (★)
Rappelons que la commande `rd.random(n)` (dans la librairie `numpy.random` avec le raccourci `rd`) permet de simuler un vecteur de taille n dont chaque coefficient est un nombre réel choisi aléatoirement entre 0 et 1.

1. Créer une série statistique contenant 10000 nombres réels choisis aléatoirement entre 1 et 5.
2. Avec `Python`, calculer la moyenne, la médiane, la variance, l'écart type et l'étendue de la série statistique.
3. Faut-il mieux regrouper cette série statistique par modalités ou par classes ? Pourquoi ?
4. Tracer l'histogramme associé à cette série statistique en la regroupant par classes (choisir 100 classes de même amplitude).
5. Tracer la boîte à moustaches associée à cette série statistique.

---

3 La librairie pandas

Commençons par importer la bibliothèque `pandas` à l'aide de la commande :

```
>>> import pandas as pd
```

Une fois cette commande exécutée, nous avons maintenant accès aux fonctions associées à cette librairie.

3.1 Exploitation des données

De nombreuses bases de données sont stockées sous la forme de fichiers numériques de format `csv` (Comma-Separated Values). Les outils de la librairie `Pandas` permettent à `Python` de les prendre en charge pour en faciliter l'analyse.

Propriété 7 (Importation de fichiers `csv`)

Si un fichier comportant des séries de données (nommé ici `fichier`) de format `csv` est chargé dans le répertoire de travail de `Python`, alors `df = pd.read_csv("fichier.csv")` affecte à la variable `df` la série de données sous la forme d'une **table** ordonnée en ligne et en colonne prête à être exploitée à l'aide des outils de la librairie `pandas`.

Exemple. On importe un fichier `csv` nommé `test` :

```
>>> df=pd.read_csv("test.csv"); print(df)

  Prenom   Age   Ville   Dispo   Permis    Enfants
0   Ines    22   Bondy       8      non          2
1    Leo    19   Paris      10      oui          0
2    Tom    20   Paris      12      non          0
3    Lea    18    Orly       8      oui          0
4   Mick    20   Paris       8      oui          0
5    Eva    22   Paris       8      oui          1
6   Mael    17   Gagny      18      non          0
```

Propriété 8 (Premières commandes)

Supposons que la variable `df` contienne une table `pandas`.

- La commande `df.head()` permet de visualiser les cinq premières lignes de la table.
- La commande `df.shape` renvoie le nombre de lignes et le nombre de colonnes de la table.

Remarque. La commande `df.head()` donne un aperçu de la table `df`. Elle est utile pour vérifier la validité des opérations commandées.

Exemple.

```
>>> df.head()

  Prenom   Age   Ville   Dispo   Permis    Enfants
0   Ines    22   Bondy       8      non          2
1    Leo    19   Paris      10      oui          0
2    Tom    20   Paris      12      non          0
3    Lea    18    Orly       8      oui          0
4   Mick    20   Paris       8      oui          0

>>> df.shape

(7, 6)
```

3.2 Indicateurs statistiques

Propriété 9 (Pour obtenir l'ensemble des indicateurs statistiques)

Si la variable `df` contient une table `pandas`, la commande `df.describe()` renvoie des informations statistiques concernant les colonnes numériques.

Exemple.

```
>>> df.describe()

                Age            Dispo          Enfants
count      7.000000         7.000000         7.000000
mean      19.714286        10.857143         0.428571
std        1.889822         3.625308         0.786796
min       17.000000         8.000000         0.000000
25%       18.500000         8.000000         0.000000
50%       20.000000        10.000000         0.000000
75%       21.000000        12.000000         0.500000
max       22.000000        18.000000         2.000000
```

On obtient dans l'ordre le nombre de lignes, la moyenne, l'écart-type, la valeur minimale, le premier, deuxième et troisième quartiles et la valeur maximale par colonne.

Remarque. Il est possible de ne visualiser qu'une colonne. Par exemple, pour n'avoir que les indicateurs de la colonne "Age", on commande : `df['Age'].describe()`.

Propriété 10 (Pour obtenir séparément les indicateurs statistiques)

Si la variable `df` contient une table `pandas`, les commandes suivantes permettent d'isoler des indicateurs statistiques :

- `df.mean()` renvoie la liste des moyennes pour chaque colonne numérique.
- `df.var()` renvoie la liste des variances pour chaque colonne numérique.
- `df.std()` renvoie la liste des écarts-types pour chaque colonne numérique.
- `df.median()` renvoie la liste des médianes pour chaque colonne numérique.
- `df.count()` renvoie le nombre de valeurs pour chaque colonne numérique.

Remarques.

1. Il est possible d'isoler la moyenne (ou autre) pour une colonne donnée. Par exemple, pour n'avoir que la moyenne de la colonne "Age", on commande : `df['Age'].mean()`.
2. Sur le même modèle, il est possible de faire la somme `sum()`, la somme cumulée `cumsum()`, le maximum `max()`, le minimum `min()` d'une colonne donnée. Par exemple, pour avoir le nombre maximum d'enfants, on commande : `df['Enfants'].max()`.

3.3 Classement, sélection

Propriété 11 (Classement et sélection conditionnelle)

Supposons que la variable `df` contienne une table `pandas`.

- La commande `df.sort_values('`*Nom_de_colonne*`')` classe la table suivant les valeurs croissantes de la colonne concernée (par ordre alphabétique dans le cas de lettres).
- La commande `df[`*condition*`]`, où *condition* est un booléen énonçant une condition partant sur `df['`*Nom_de_colonne*`']`, affiche les lignes dont la condition est réalisée.

Remarques.

1. Pour définir la *condition*, on pourra utiliser les comparaisons `==, <, >, <=, >=, !=` et les connecteurs logiques `&` (et) et `|` (ou).
2. On peut également obtenir un classement par ordre décroissant des valeurs d'une colonne avec la commande `df.sort_values('`*Nom_de_colonne*`',ascending==False)`.

Exemple. Reprenons la table `df` utilisée précédemment.

- Pour classer la table par ordre croissant suivant les ages :

```
>>> df.sort_values('Age')
```

	Prenom	Age	Ville	Dispo	Permis	Enfants
6	Mael	17	Gagny	18	non	0
3	Lea	18	Orly	8	oui	0
1	Leo	19	Paris	10	oui	0
2	Tom	20	Paris	12	non	0
4	Mick	20	Paris	8	oui	0
0	Ines	22	Bondy	8	non	2
5	Eva	22	Paris	8	oui	1

- Pour sélectionner les personnes qui ont 20 ans :

```
>>>df[df['Age']==20]
```

	Prenom	Age	Ville	Dispo	Permis	Enfants
2	Tom	20	Paris	12	non	0
4	Mick	20	Paris	8	oui	0

- Pour sélectionner les personnes disponibles 8 heures et qui habitent Paris :

```
>>> df[(df['Dispo']==8) & (df['Ville']=='Paris')]
```

	Prenom	Age	Ville	Dispo	Permis	Enfants
4	Mick	20	Paris	8	oui	0
5	Eva	22	Paris	8	oui	1

Exercice 3 (★)

On s'intéresse à certains aspects des 20 villes les plus peuplées de France. On s'intéresse plus particulièrement :

- à leur région d'appartenance ;
- à leur population (en nombre d'habitant) ;
- à leur densité (en habitants par km^2 ;
- à leur taux de chômage (en pourcentage de la population) ;
- à leur label "ville fleurie" ;
- à leur label "ville connectée" (en nombre de @) ;
- à leur pluviométrie (en mm) ;
- à leur ensoleillement (en heure).

Ces données sont saisies dans un fichier nommé `TP3-ex3` converti au format `csv` disponible sur mon site `anthony-mansuy.fr/`. Écrire les commandes qui permettent de :

1. Importer le fichier `TP3-ex3` et l'afficher sur `Python`.
2. Donner les indications statistiques correspondant à la population.
3. Déterminer la moyenne du taux de chômage pour ces 20 villes.
4. Classer les villes par ordre croissant de leur indice pluviométrique.
5. Isoler les villes (s'il y en a) dont la densité est inférieure à 4000 hab/km^2 et dont la durée annuelle d'ensoleillement est d'au moins 2000 heures.
6. Isoler les villes (s'il y en a) de la région PACA et classées "4 fleurs".

Exercice 4 (★★)

À partir de mon site `anthony-mansuy.fr/`, télécharger le fichier `TP3-ex4` converti au format `csv`. C'est un tableau dont les colonnes sont :

- `Index`, qui contient l'index des pays ;
- `Pays`, qui contient les noms des pays ;
- `Superficie`, qui contient la surface terrestre en milliers de km^2 de chaque pays ;
- `Population`, qui contient le nombre d'habitants en millions de chaque pays ;
- `Naissance`, qui contient le nombre de naissances sur 1000 habitants ;
- `Mort`, qui contient le nombre de décès sur 1000 habitants ;
- `Homme`, qui contient l'espérance de vie des hommes ;
- `Femme`, qui contient l'espérance de vie des femmes.

Ces données sont issues de l'étude 2017 de l'Institut National d'Études Démographiques (disponible également sur mon site). Pour répondre aux questions qui suivent, vous trouverez en annexe de ce TP la liste des pays et leurs index.

1. Importer et afficher le fichier `TP3-ex4` sur `Python`.
2. (a) Calculer la surface terrestre mondiale, le nombre d'habitants mondial et la densité moyenne d'habitants au km^2.
   (b) Calculer la surface terrestre, le nombre d'habitants et la densité moyenne d'habitants au km^2 pour chaque continent.
   (c) Représenter la densité moyenne d'habitants au km^2 pour chaque continent en utilisant un diagramme en bâtons (on mettra en abscisse des entiers de 1 à 6).

3. On considère l'espérance de vie des hommes et des femmes par pays.
   (a) Calculer la moyenne sur l'ensemble des pays.
   Ce résultat correspond-il à l'espérance de vie mondiale des hommes et des femmes ?
   (b) Calculer la médiane, la variance et l'écart-type.
   (c) Représenter l'histogramme de l'espérance de vie des hommes sur l'intervalle $[0, 100]$ avec 20 classes. Quelle est la classe modale de l'espérance de vie des hommes ?
   (d) Classer la table suivant les valeurs croissantes de la colonne `Femme`.
   En déduire le pays où l'espérance de vie des femmes est la plus grande et celui où elle est la plus petite.
   (e) A l'aide de la commande `describe`, déterminer les valeurs du premier et du troisième quartile ainsi que l'écart inter-quartile de la colonne `Femme`.
   En déduire la liste des pays dont l'espérance de vie est inférieure au premier quartile.
4. On rappelle que le taux d'accroissement naturel est la différence entre la natalité et la mortalité.
   (a) Quels sont les accroissements minimaux et maximaux ?
   (b) Faire afficher la liste des pays pour lesquels l'accroissement est négatif.
   (c) Déterminer l'accroissement mondial moyen.
   (d) Dans ses projections, l'INED prévoit une population mondiale de 9731 millions d'habitants en 2050. Cela est-il conforme à l'hypothèse d'un taux d'accroissement constant ?

Annexe : Liste des pays et de leurs index

Afrique

Afrique septentrionale

1. Algérie
2. Égypte
3. Libye
4. Maroc
5. Sahara occidental
6. Soudan
7. Tunisie

Afrique occidentale

8. Bénin
9. Burkina Faso
10. Cap-Vert
11. Côte d'Ivoire
12. Gambie
13. Ghana
14. Guinée
15. Guinée-Bissau
16. Liberia
17. Mali
18. Mauritanie
19. Niger
20. Nigeria
21. Sénégal
22. Sierra Leone
23. Togo

Afrique orientale

24. Burundi
25. Comores
26. Djibouti
27. Érythrée
28. Éthiopie
29. Kenya
30. Madagascar
31. Malawi
32. Maurice
33. Mayotte
34. Mozambique
35. Ouganda
36. Réunion
37. Rwanda
38. Seychelles
39. Somalie
40. Sud-Soudan
41. Tanzanie
42. Zambie
43. Zimbabwe

Afrique centrale

44. Angola
45. Cameroun
46. Centrafricaine (Rép.)
47. Congo
48. Congo(Rép. dém.)
49. Gabon
50. Guinée équatoriale
51. Sao Tomé-et-Principe
52. Tchad

Afrique australe

53. Afrique du Sud
54. Botswana
55. Lesotho
56. Namibie
57. Swaziland

Amérique du Nord

Amérique septentrionale

58. Canada
59. États Unis

Amérique centrale

60. Belize
61. Costa Rica
62. Guatemala
63. Honduras
64. Mexique
65. Nicaragua
66. Panama
67. Salvador

Caraïbes

68. Antigua-et-Barbuda
69. Aruba
70. Bahamas
71. Barbade
72. Cuba
73. Curaçao
74. Dominicaine (Rép.)
75. Dominique
76. Grenade
77. Guadeloupe
78. Haïti
79. Jamaïque
80. Martinique
81. Porto Rico
82. Sainte Lucie
83. St Vincent Grenadines
84. St.Kitts-et-Nevis
85. Trinité-et-Tobago
86. Vierges (Îles)

Amérique du Sud

87. Argentine
88. Bolivie
89. Brésil
90. Chili
91. Colombie
92. Équateur
93. Guyana
94. Guyane (française)
95. Paraguay
96. Pérou
97. Surinam
98. Uruguay
99. Venezuela

Asie

Asie occidentale

100. Arabie saoudite
101. Arménie
102. Azerbaïdjan
103. Bahreïn
104. Chypre
105. Émirats arabes unis
106. Géorgie
107. Irak
108. Israël
109. Jordanie
110. Koweït
111. Liban
112. Oman
113. Palestine (Territoires)
114. Qatar

115. Syrie
116. Turquie
117. Yémen

Asie centrale

118. Kazakhstan
119. Kirghizistan
120. Tadjikistan
121. Turkménistan
122. Ouzbékistan

Asie du sud

123. Afghanistan
124. Bangladesh
125. Bhoutan
126. Pakistan
127. Inde
128. Iran
129. Maldives
130. Népal
131. Sri Lanka

Asie du sud-ouest

132. Brunei
133. Cambodge
134. Indonésie
135. Laos
136. Malaisie
137. Myanmar (Birmanie)
138. Philippines
139. Singapour
140. Thaïlande
141. Timor-Est
142. Viêt Nam

Asie orientale

143. Chine
144. Chine-Hong Kong
145. Chine-Macao
146. Corée du Nord
147. Corée du Sud
148. Japon
149. Mongolie
150. Taïwan

Europe

Europe septentrionale

151. Danemark
152. Estonie
153. Finlande
154. Irlande
155. Islande
156. Lettonie
157. Lituanie
158. Norvège
159. Royaume-Uni
160. Suède

Europe occidentale

161. Allemagne
162. Autriche
163. Belgique
164. France (métropolitaine)
165. Liechtenstein
166. Luxembourg
167. Monaco
168. Pays-Bas
169. Suisse

Europe orientale

170. Biélorussie
171. Bulgarie
172. Hongrie
173. Moldavie
174. Pologne
175. Roumanie
176. Russie
177. Slovaquie
178. Tchèque (République)
179. Ukraine

Europe méridionale

180. Albanie
181. Andorre
182. Bosnie-Herzégovine
183. Croatie
184. Espagne
185. Grèce
186. Italie
187. Kosovo
188. Macédoine
189. Malte
190. Monténégro
191. Portugal
192. Saint-Marin
193. Serbie
194. Slovénie

Océanie

195. Australie
196. Fidji
197. Guam
198. Kiribati
199. Marshall (Îles)
200. Micronésie (États fédérés de)
201. Nouvelle-Calédonie
202. Nouvelle-Zélande
203. Papouasie-Nouvelle Guinée
204. Polynésie française
205. Salomon (Îles)
206. Samoa occidentales
207. Tonga
208. Vanuatu