3 T.D. statistiques descriptives

Les objectifs : Description d'une série statistique : effectifs, fréquences, fréquences cumulées. Représentations graphiques.
Caractéristiques de position (moyenne, médiane, mode) et de dispersion (variance s_x^2 et écart-type s_x, quartiles, déciles)
Séries statistique double de taille n portant sur deux caractères quantitatifs x et y. Point moyen $(\bar{x}, \bar{y})$ du nuage de points de $\mathbb{R}^2$ associé.
Caractéristiques d'une série statistique double (covariance s_{xy}, coefficient de corrélation r_{xy}, ajustement affine selon la méthode des moindres carrés ou régression linéaire). Interprétation géométrique de l'ajustement affine.

Exercice 1 :

On considère une série statistique de 50 mesures d'envergure d'ailes déployées (en cm) mesurées chez des vautours adultes de la région de l'Adour. La série est rangée par valeurs croissantes avec sur la première ligne les femelles et sur la seconde les mâles.

Femelles	103	110	111	112	118	120	125	125	126	127	128	130
	132	139	139	139	142	145	148	150	151	153	156	160
Males	141	144	146	148	150	151	153	155	156	161	163	163
	165	168	169	170	171	172	175	175	176	177	178	179

① Créer sous Python une liste de liste appelée `tab` et qui représente ces données. La transformer en un tableau avec lequel on puisse travailler avec la bibliothèque « numpy ».

② Quelles informations peut-on obtenir, sans calcul, en examinant les données ?

③ Calculer la moyenne et la variance des distributions des femelles et des mâles.

④ Déterminer l'intervalle interquartile de la série des femelles et de la série des mâles.

⑤ Afin de simplifier les données qui sont supposées continues, procéder à un regroupement en classes de largeur 10. Donner pour chacune des séries, le centre des classes, les effectifs et fréquences de chaque classe. En assimilant chaque classe à son centre, calculer la moyenne et la variance des données regroupées et comparer aux résultats obtenus en 3.

Exercice 2 :

Ma fille, cet été, m'a fait remarquer après avoir inlassablement effeuillé des marguerites, qu'en récitant la comptine « je t'aime, un peu, beaucoup, passionnément, à la folie, pas du tout », elle semblait s'arrêter plus souvent sur « un peu » que sur les autres...
Vous trouverez dans le fichier `marguerites.csv` cinq séries de cinquante données dénombrant les pétales de pâquerettes prises dans cinq champs différents.

① Calculer la moyenne et la variance de chacune de ces séries. Représentez graphiquement les moyennes du nombre de pétales des cinq séries.

② Représenter grâce à une boite à moustache ces cinq séries de données. En quoi cela justifie-t-il qu'on travaille ensuite l'ensemble des 250 données.

③ Écrire une fonction Python permettant de retourner la fréquence des issues de chacune des comptines réalisées à partir des 250 fleurs échantillonnées. Qu'en concluez-vous ?

Exercice 3 :

Dans le cadre d'une étude portant sur les dépenses mensuelles moyennes (alimentaires/habillement-loisirs) des ménages français composés d'un couple et d'un enfant mineur avant le passage à l'euro, les données suivantes ont été retrouvées. *unité :* 1 F.

Revenus mensuel moyen x_i	3980	6030	7940	10000	15140	25120	39810
Dépenses alimentaires y_i	3090	3715	4170	4570	5510	7080	8710
Dépenses habillement-loisirs z_i	200	407	661	1000	2042	4786	10223

① Représenter graphiquement sur une même première figure le nuage des points $M_i(x_i, y_i)$ et des points $N_i(x_i, z_i)$.
Faire de même dans une deuxième figure avec une double échelle logarithmique. Qu'en concluez-vous ?

② On pose $u_i = \log(x_i)$, $v_i = \log(y_i)$ et $w_i = \log(z_i)$.

a. Écrire une fonction Python `covariance(X,Y)` et une fonction `coeffCorr(X,Y)` retournant la covariance et le coefficient de corrélation de deux séries de données X et Y.
Vérifier alors la validité de l'ajustement linéaire du nuage précédent en calculant le coefficient de corrélation linéaire entre U et V et entre U et W.

b. Écrire une fonction Python `regression(X,Y)` permettant à la fois d'obtenir les équations des droites de régression de V en U et de W en U et de tracer leur représentation graphique sur les nuages de points.

c. En déduire des relations de la forme : $Y = k \cdot X^{\alpha}$ et $Z = \dfrac{h}{10^5} \cdot X^{\beta}$.
✐ On arrondira k, h, α et β à 10^{-2} près.

③ **a.** Déterminer les dépenses alimentaires et les dépenses d'habillement-loisirs moyennes d'un ménage dont le revenu mensuel moyen est de 20000 F.

b. Déterminer le revenu mensuel d'un ménage qui dépense autant pour la nourriture que pour l'habillement-loisirs.

Exercice 4 : Cinétique chimique

On donne les concentrations d'un réactif en solution à différents instants :

Temps / s	0	60	108	162	230	326	363	449	527	612	709	822	947
c(t) / $mmol.L^{-1}$	5,00	4,73	4,46	4,19	3,92	3,66	3.4	3,14	2,88	2,63	2,38	2,13	1,89

Temps / s	1088	1234	1678	1974
c(t) / $mmol.L^{-1}$	1,64	1,40	1,16	0,93

On suppose qu'on a une évolution vérifiant une équation différentielle du type $c'(t) = -kc^n(t)$ [*]. On cherche à déterminer l'ordre n de la réaction et la constante k de vitesse de réaction.

① Méthode différentielle :

- **a.** Écrire une fonction Python permettant d'évaluer les vitesses de réactions à partir des données ci-dessus. Tracer dans deux fenêtres distinctes la courbe représentant c en fonction de T et celle représentant c' en fonction de T.
- **b.** Expliquer pourquoi le tracé de $\ln(-c')$ en fonction de $\ln(c)$ permet d'estimer l'ordre de la réaction.
- **c.** Proposer une fonction Python `dteRegression()` qui, pour X et Y donnés en entrée, renvoie les coefficients a et b de la droite de régression de la série statistique (X, Y). En déduire deux valeurs possibles pour l'ordre n de la réaction.

② Méthode d'intégration :

- **a.** Intégrer l'équation différentielle ci-dessus pour les valeurs de n envisagées à l'issue de la méthode différentielle.
- **b.** Proposer dans chaque cas un changement de variable qui permette de décider quel est l'ordre de la réaction.
- **c.** Déterminer la constante k de vitesse de réaction.

On considère une série statistique S de taille n portant sur un caractère x. Les valeurs observées x_1,...,x_n seront considérées comme des réalisations de variables aléatoires mutuellement indépendantes X_1,...,X_n ayant toutes la même loi qu'une variable aléatoire « abstraite »X appelée *variable mère*. Choisir ce modèle suppose que le phénomène étudié soit bien défini, invariant au cours des observations et que ces observations n'exercent aucune influence entre elles.

➤ **Caractéristiques de position** :

→ **moyenne** (`numpy.mean()`) : $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$

→ **mode** : observation x_k dont l'effectif n_k est maximum.

→ **médiane** (`numpy.median()`) : On suppose les observations classées par ordre croissant sous la forme : $x_{(1)}$,...,$x_{(n)}$.
On appelle médiane l'observation $x_{(k)}$ telle que la moitié des observations lui sont inférieure. Deux cas se présentent :

- Si moins de 30 données : On distinguera le cas pair du cas impair. Si le nombre n de données est impair, il suffit de prendre pour la médiane $x_{(k)}$ où $k = \dfrac{n+1}{2}$ et si n est pair (on posera $n = 2p$) la médiane est choisie comme la demi-somme de la p-ième valeur et de la $(p+1)$-ième valeur, à savoir : $x_{(k)} = \dfrac{x_{(p)} + x_{(p+1)}}{2}$.
  ✎ Pour certains auteurs, il suffit de prendre $k = \lceil n/2 \rceil$ où $\lceil x \rceil$ est le premier entier n tel que $x \leq n$.
- Si plus de 30 données : On utilisera de préférence le tableau de **fréquences cumulées** (`cumsum()`) qui autorise une recherche aussi bien graphique que calculatoire de la médiane (antécédent de $n/2$).

➤ **Caractéristique de dispersion** :

→ **variance** (`numpy.var()`) : $\sigma_x^2 = \sum_{i=1}^{p} f_i(x_i - \bar{x})^2 = \sum_{i=1}^{p} f_i x_i^2 - \bar{x}^2 = \overline{x^2} - \bar{x}^2$

→ **quantile d'ordre** α (`numpy.percentile()`) (« quartiles » si $\alpha = 1/4$ ou $\alpha = 3/4$ et déciles pour $\alpha = i/10$, $i \in [\![1, 9]\!]$) est l'observation $x_{(k)}$ où $k = \lceil \alpha n \rceil$.

→ Graphiques : `matplotlib.pyplot.hist()` et `matplotlib.pyplot.boxplot()`

➤ **Séries multivariées** :

→ **covariance** (`numpy.cov(x,y,bias = 1)`) :

$$s_{x,y} = \begin{cases} \dfrac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x} \cdot \overline{y} \text{ si variable non groupées} \\ \dfrac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{i,j} x_i y_j - \overline{x} \cdot \overline{y} \text{ si variable groupées} \end{cases}$$

→ **Coefficient de corrélation** (`numpy.corrcoef()`) : $r_{x,y} = \dfrac{s_{x,y}}{s_x s_y}$

→ **Point moyen** : $G(\overline{x}, \overline{y})$

→ **Régression linéaire** (`(a,b)=polyfit(x,y,1),plot(x,y,'o',x,a*x+b,'-')`) :

$$(\Delta) : y - \overline{y} = \frac{s_{x,y}}{s_x^2}(x - \overline{x}) \text{ ou encore } y = ax + b,\ a = \frac{s_{x,y}}{s_x^2} \text{ et } b = \overline{y} - a \cdot \overline{x}$$