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 procède à l'analyse chimique de massifs granitiques de la chaîne des Pyrénées dont nous retenons les teneurs en silice (pourcentage pondéral arrondi) :

Teneur	67	65	70	72	72	71	72	75	71	74
Teneur	76	76	76	78	75	76	74	72	76	75

① Écrire une fonction `moyenne` qui prend en entrée une liste `X` (non vide) de nombres réels et retourne la moyenne des éléments de la liste.

② Écrire une fonction `variance` qui prend en entrée une liste `X` (non vide) de nombres réels et retourne la variance des éléments de la liste.

③ Calcul la moyenne $\overline{x}$ et l'écart-type s de cette série de $n = 20$ observations. Préciser en particulier les valeurs des trois quartiles empiriques $q_{0,25}$, $q_{0,5}$ et $q_{0.75}$ ainsi que le mode. En fournir une représentation sous la forme d'une boîte à moustaches.

④ La variable statistique étudiée (teneur en silice) étant considérée comme discrète, tracer l'histogramme des fréquences.

⑤ La variable statistique est maintenant considérée comme continue (les égalités entre certaines valeurs s'interprétant comme une conséquence des arrondis). On regroupe les données en classes d'amplitude 2.

a. Donner sous forme d'un tableau le centre, l'effectif de chaque classe, ainsi que les fréquences cumulées.
b. Calculer à nouveau la moyenne de cette nouvelle distribution. Comparer ce résultat à celui obtenu sur les données brutes.
c. Tracer la courbe des fréquences empiriques cumulées et proposer un calcul approché de la médiane.

Exercice 3 ✶ :

L'objectif est de travailler sur l'évolution des débits de la Loire au cours des années 1989 à 2007. La DIREN a fourni au format « .CSV » le tableau de 6758 données intitulé « debitLoire1.csv » et disponible sur le site de la classe.

① Fournir une description statistique des deux séries fournies accompagnée d'un histogramme.

② Donner une représentation chronologique des valeurs de la seconde série sur 90 jours consécutifs à partir du 1er janvier 2005. Interpréter.

Exercice 4 ✶ :

On a répertorié le nombre de frères et soeurs de tous les étudiants de sciences inscrits en L1 à l'université de Nantes.

Fratrie	0	1	2	3	4	**Total**
Effectif	38	94	75	48	5	260

① Calculer la moyenne et la médiane de cette distribution.

② Six étudiants absents lors du recueil des données annoncent qu'ils ont respectivement 2, 3, 2, 0, 4 et 2 frères et soeurs. Calculer la moyenne et la médiane de la série une fois complétée et les comparer aux valeurs précédentes.

Exercice 5 ✶ Longueurs des oeufs de coucou

Nous disposons grâce au site « Data and Story Library » de données sur la longueur des oeufs de coucou déposés dans les nids d'autres oiseaux. Toutes les longueurs sont en mm.

Pipit des prés	19.65	20.05	20.65	20.85	21.65	21.65	21.65	21.85	21.85	21.85
	22.05	22.05	22.05	22.05	22.05	22.05	22.05	22.05	22.05	22.05
	22.25	22.25	22.25	22.25	22.25	22.25	22.25	22.25	22.45	22.45
	22.45	22.65	22.65	22.85	22.85	22.85	23.05	23.25	23.25	23.45
	23.65	23.85	24.25	2445						
Pipit des bois	21.05	21.85	22.05	22.45	22.65	23.25	23.25	23.25	23.45	23.45
	23.65	23.85	24.05	24.05	24.05					
Roitelet	19.85	20.05	20.25	20.85	20.85	20.85	21.05	21.05	21.05	21.25
	21.45	22.05	22.05	22.05	22.25					

① Fournir une description de la série portant sur la taille des oeufs dans les nids de Pipit des prés.

② Il y a eu une erreur de saisie. Corriger-la et mesurer son impact sur la moyenne et la médiane.

③ Existe-t-il selon vous une différence significative de la taille des oeufs selon l'hôte choisi ?

Exercice 6 ✶✶ : régression

En écologie, il est possible de mettre en évidence la relation existant entre le nombre N d'espèces présentes dans un habitat donné (bien délimité) et la surface S de cet habitat. L'une des lois couramment utilisée assure que :

$$N = AS^B,\ A, B \in \mathbb{R}$$

Afin de vérifier la validité de cette relation pour les plantes présente dans une prairie (pissenlit, pâquerettes, orties, boutons d'or, etc.) on a effectué les mesures indiquées dans le tableau ci-dessous.

S	1	2	3	4	8	12	16	32	64	128
N	6	6	7	8	9	10	11	13	15	15

① Représenter les valeurs de N en fonction de celles de S.

② Ecrire une fonction `covariance(X,Y)` qui retourne la covariance de deux listes `X` et `Y` si elle existe et le booléen `False` sinon.

③ Proposer un changement de variable qui rende possible une régression linéaire. Écrire une fonction `coeffcorr(X,Y)` qui retourne le coefficient de corrélation de deux listes `X` et `Y` s'il existet. En déduire le coefficient de corrélation obtenu après changement de variable.

④ Écrire une fonction `regression(X,Y)` qui trace la droite de régression d'équation $y = ax + b$ et retourne ses coefficients a et b.
Proposer des valeurs possibles pour A et B dans le modèle proposé. Confrontez pour ces valeurs le graphe de $N = AS^B$ avec celui des données expérimentales.

⑤ Quelle valeur $\tilde{N}$ ce modèle prédit-il pour le nombre d'espèces pouvant coexister dans un habitat de surface $S = 100$?

Exercice 7 ✶✶ : régression

On a mesuré sur un peuplement de bouleau blanc (*Betula alba*) dans le Massif Central les circonférences des troncs de 21 individus à la hauteur de 1.3 mètres du sol (indice DBH). Dans le même temps, un carottage des arbres a permis d'estimer leurs ages respectifs. De cet ensemble de données on a extrait les données des arbres d'ages de 1 à 120 par pas de 20 ans.
Par ailleurs, on a constaté sur le terrain que les arbres se répartissent en trois catégories : les arbres les plus hauts (dominants), les arbres moyens (codominants) et les arbres plus petits, sous le couvert des autres (les dominés).

Ages	1	20	40	60	80	100	120
Dominants	1.26	22.29	40.09	56.15	63.49	71.69	81.08
Dominés	1.27	16.02	29.42	31.61	35.61	35.69	35.93
Codominants	1.29	22.14	35.69	49.23	56.88	60.43	63.74

① Tracer sur un même graphique les trois courbes représentant la circonférence des troncs en fonction de l'age. Que constate-t-on ? Quel type de fonction peut-on envisager d'ajuster ?

② On souhaite vérifier que la croissance en circonférence des troncs peut être modélisée par une « exponentielle saturée » de la forme :

$$y(t) = y_m(1 - e^{-rt})$$

où $y(t)$ est la circonférence à l'instant t, $r \in \mathbb{R}_+^*$ est le taux de croissance en circonférence et t est le temps.

a. Comment interprétez vous y_m ? Les valeurs estimées expérimentalement sont respectivement de 91.2 cm, 65.43 cm et 36.00 cm pour chacune des trois catégories d'arbres.

b. En remarquant que, d'après l'expression de $y(t)$, la quantité $\ln(y_m - y(t))$ dépend de façon linéaire de t, estimer au moyen d'une régression linéaire le paramètre r pour chacun de ces trois modèles.

c. Vérifier sur l'un des trois résultats la bonne qualité de l'ajustement des données.

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()`) : $s_x^2 = \sum_{i=1}^{p} f_i(x_i - \bar{x})^2 = \sum_{i=1}^{p} f_i x_i^2 - \bar{x}^2 = \overline{x^2} - \bar{x}^2$

→ **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,'-')`) :

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