Informatique - TP 13

Un peu de statistiques

M. Marmorat, M. Morel

13 Juin 2024

Pour ce TP, on téléchargera le fichier `TP13-données.py` *depuis cahier de prépa » Informatique » TP » Données. Attention, pour toutes les fonctions, il faudra d'abord écrire une fonction prenant en argument des variables quelconques, puis les tester avec les données proposées dans ce fichier.*

Exercice 1 Moyenne et médiane

On appelle série statistique à valeurs réelles toute famille finie de nombres réels : $x_1, x_2, \dots, x_N$. L'entier $N \in \mathbb{N}$ est appelé effectif de la série. En Python, on représentera une série statistique par une liste.

La moyenne arithmétique moy de cette série est alors donnée par la formule : $moy = \dfrac{1}{N}\displaystyle\sum_{k=1}^{N} x_k$.

Q1 Écrire une fonction calculant la moyenne arithmétique d'une série.

La médiane med de cette série est une valeur séparant ces données en deux groupes de même taille : celui des éléments inférieurs à med, et celui des éléments supérieurs à med. En d'autres termes, med est une médiane de la série $x_1, \dots, x_N$ lorsque 50% des x_i sont plus grands que med et 50% sont plus petits.

Par exemple pour une série représentée par la liste Python `[1,6,4,10,12,3,7]` alors la médiane est l'élément $\boxed{med =}$ En effet, il y a autant d'éléments de la liste strictement inférieurs à med que strictement supérieurs à med.

Pour calculer la médiane d'une série de nombres, on commence par trier ces nombres dans l'ordre croissant. Si on note N l'effectif total de la série, et $y_1, \dots, y_N$ la série triée dans l'ordre croissant, alors on pourra prendre pour médiane med :

- si N est impair : l'élément au milieu de la serie triée, c'est-à-dire $\boxed{med =}$
- si N est pair : la moyenne entre les deux éléments au milieu de la série triée, c'est-à-dire $\boxed{med =}$

Q2 Écrire une fonction calculant la médiane d'une série. On utilisera la commande `L.sort()` qui modifie une liste de nombres `L` en la liste triée dans l'ordre croissant (même si on sait déjà le faire grâce à un tri par sélection ou un tri par insertion...).

Q3 Avez-vous testé votre fonction à la question précédente ? (cela doit être un réflexe).

Q4 Donner un exemple de votre choix de liste à 5 éléments dont :

- la moyenne est strictement supérieure à la médiane,
- la médiane est strictement supérieure à la moyenne.

Pour évaluer l'inflation, on réalise l'expérience suivante dans un supermarché : au 1er janvier 2022, 30 clients du magasin sont sélectionnés et notent le montant total de leurs courses. Au 1er janvier 2023, on demande à chacun de ces 30 clients de racheter exactement les mêmes produits, et on note à nouveau les prix payés en caisse.

Le fichier `TP13-donnees.py` donne une matrice `prix` comportant 2 lignes et 30 colonnes. Chaque colonne correspond à un client, la première ligne correspond au prix payé en 2022, et la deuxième à celui payé en 2023.

Q5 Écrire une fonction prenant en argument une matrice de 2 lignes similaire à la matrice `prix` et renvoyant une liste donnant le pourcentage d'augmentation du prix payé par chaque client. Par exemple, pour la matrice $\begin{pmatrix} 10 & 50 & 100 \\ 15 & 60 & 98 \end{pmatrix}$, votre fonction doit renvoyer la liste `[50, 20, -2]`.

Q6 Une fois l'étude terminée, le gérant du magasin annonce : *«notre politique de prix a permis de contenir l'inflation sous la barre des* 10%*, puisqu'en moyenne les clients concernés par l'étude ont vu le prix de leur courses augmenter de* $8,5\%$*»*. Interrogés, les clients n'ont pas du tout la même impression. Confirmez-vous le chiffre donné par le gérant ? Comment expliquez-vous l'impression des clients ?

Exercice 2 Régression linéaire

On appelle série statistique bivariée à valeurs réelles toute famille de couples de nombres réels : $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$. En Python, on représentera une telle série par 2 listes (une pour la série des x_i et une pour celle des y_i).

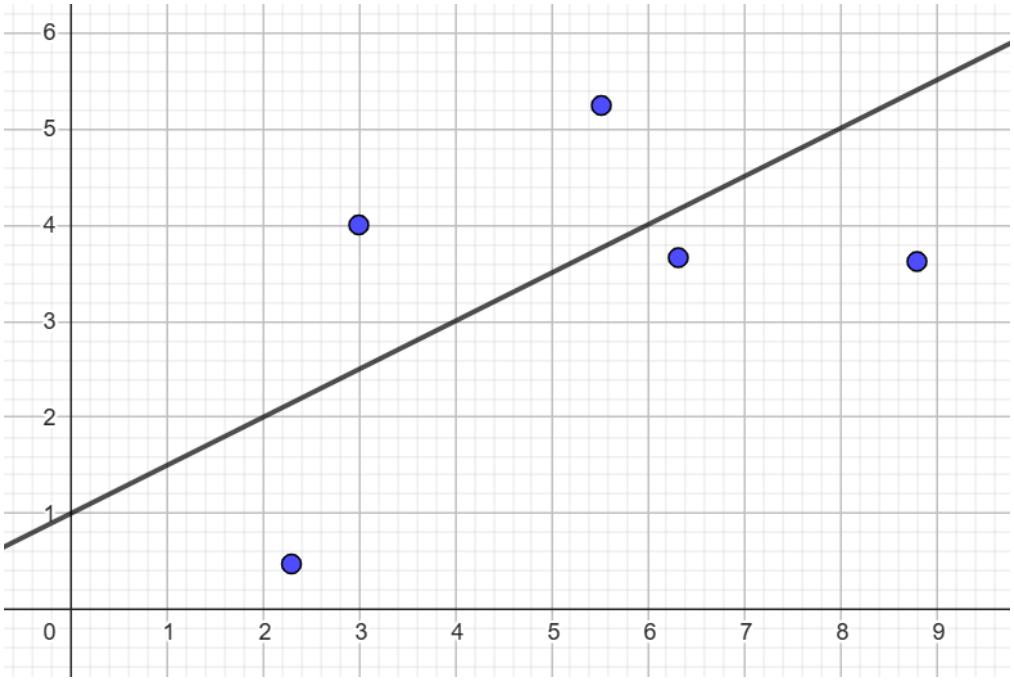
Si on suspecte que la quantité mesurée en y_i dépend de celle mesurée en x_i, il est judicieux de tracer un graphe où on place les valeurs x_i en abscisses et les valeurs y_i en ordonnées. On obtient alors un nuage de points.

Q1 Récupérer les listes `list_x` et `list_y` du fichier `TP13-données` et tracez le nuage de points correspondant.

Si on suspecte que la quantité mesurée en y_i dépend de manière affine de celle mesurée en x_i (c'est-à-dire que : $\exists\, a, b \in \mathbb{R} \ : \ \forall\, i,\ y_i = ax_i + b$), il est judicieux de tracer une droite de régression linéaire. La méthode de régression linéaire des moindres carrés consiste à trouver cette droite, d'équation $y = ax + b$, en minimisant la quantité

$$f(a,b) = \sum_{k=1}^{N} \Big| y_i - (ax_i + b) \Big|^2.$$

Q2 Indiquer sur le schéma ci-dessous représentant un nuage de points (x_i, y_i) et une droite d'équation $y = ax+b$ à quoi correspond la quantité $f(a,b)$.



On peut démontrer[1] que les valeurs a et b minimisant la quantité $f(a,b)$ sont données par les formules suivantes :

$$a = \frac{N\sum_{k=1}^{N} x_k y_k - \sum_{k=1}^{N} x_k \sum_{k=1}^{N} y_k}{N\sum_{k=1}^{N} x_k^2 - \left(\sum_{k=1}^{N} x_k\right)^2} \quad \text{et} \quad b = \frac{1}{N}\left(\sum_{k=1}^{N} y_k - a x_k\right).$$

Q3 Écrire une fonction prenant en arguments 2 listes $[x_1, \ldots, x_N]$ et $[y_1, \ldots, y_N]$ et renvoyant les valeurs a et b correspondantes. On emploiera le moins de boucles `for` possibles et on ne demandera pas à Python de recalculer plusieurs fois la même quantité.

Q4 Tracer la droite de régression linéaire pour le schéma de la question 1.

Si on suspecte que la quantité mesurée en y_i dépend de x_i d'une manière autre qu'affine, il est possible dans certains cas simples de se ramener à tracer une droite de régression linéaire. Il faut pour cela changer de variables et tracer la droite de régression linéaire entre les quantités $u(x_i)$ et $v(y_i)$ où u et v sont des fonctions simples bien choisies, généralement des puissances ou des logarithmes.

Plus une région est vaste, plus le nombre d'espèces y vivant est grand. Pour modéliser mathématiquement ce phénomène (et mesurer ainsi la biodiversité), les scientifiques utilisent parfois la loi SPAR ("species-area relationship"). Elle stipule que si S représente la surface du terrain étudié et E le nombre d'espèces présentes sur ce terrain, alors on a :

$$E = c\,S^d$$

où c et d sont des constantes à ajuster selon la région étudiée.

On récolte des plantes dans des prairies de la région. Les données récoltées sont résumées dans les deux listes du fichier `TP13-donnes.py` : la liste `list_E` donne le nombre d'espèces rencontrées et la liste `list_S` la surface explorée pour dénombrer ces espèces (en m^2).

Q5 *(Sans ordinateur)*. Montrer que la loi de SPAR revient à établir une relation affine entre $\ln(E)$ et $\ln(S)$.

Q6 Tracer le nuage de points représentant $\ln(E)$ en fonction de $\ln(S)$.

Q7 Calculer et tracer la droite de régression linéaire associée.

1. voir un exercice dans le chapitre sur les fonctions de 2 variables

Q8 En déduire les constantes c et d adaptées aux données mesurées.

Q9 Ce modèle est-il réaliste pour une surface S grande ?

Exercice 3 BPCRST

Récupérez dans le fichier `TP13-données` la liste des nombres de cas quotidiens de covid-19 déclarés en France durant approximativement les 2 première années de l'épidémie.

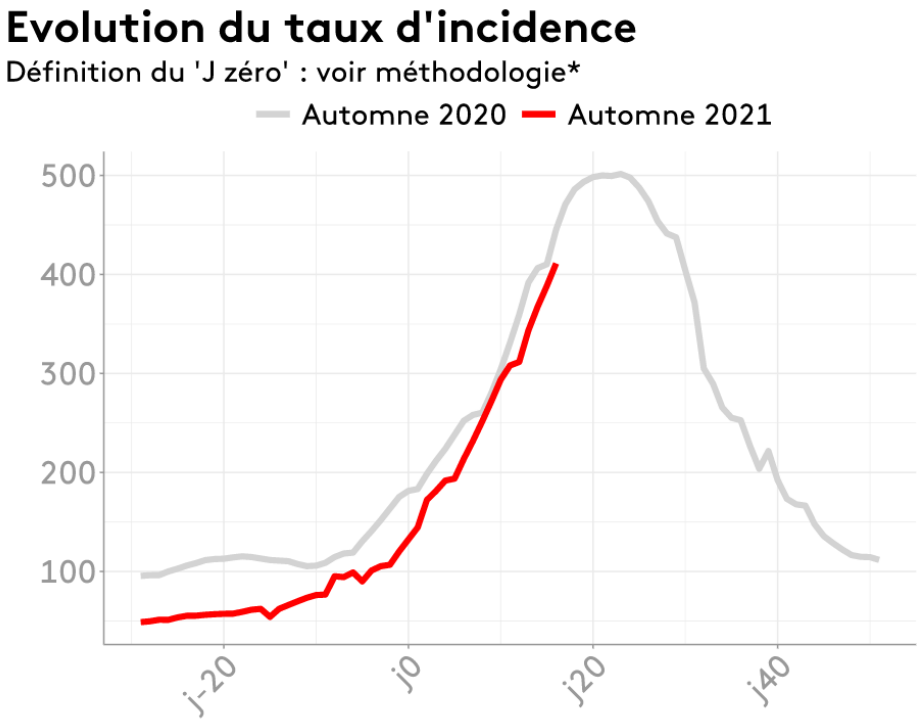
Q1 Tracer le graphe du nombre de cas en fonction du temps.

Le graphe précédent présente de nombreux "pics" liés aux weekends, pendant lesquels les données sont moins remontées. Pour lisser ces irrégularités, on représente souvent la *moyenne* du nombre de cas sur une semaine glissante.

Q2 Écrire une fonction prenant en argument une liste de nombres et renvoyant la liste des moyennes glissantes de ces nombres sur une plage de longueur 7.Si la liste prise en argument est de longueur N, alors la liste renvoyée sera donc de longueur $N-7$.

Q3 Créer alors la liste `cas_moy` contenant les nombres moyens de cas de Covid-19 lissés sur 7 jours puis représenter l'évolution du nombre de cas. Vérifier que la liste `cas_moy` commence bien par `[27.28, 28.42, 38.14, 55.0...]`.

Dans un article de presse (suivre ce lien), des journalistes souhaitent comparer la vague de l'automne 2020 à celle de l'automne 2021. Ils souhaitent pour cela représenter sur le même graphique l'évolution du nombre de cas depuis le début de chaque vague et obtiennent le graphique suivant [2] :



On souhaite reproduire ce graphique avec nos données, *on utilisera la liste* `cas_moy` *du nombre moyen de cas* (à la place du taux d'incidence utilisé dans l'article).

On donne les repères chronologiques suivants dans la liste de cas fournie :

- l'automne 2020 se situe approximativement entre les indices 200 et 300,
- l'automne 2021 se situe approximativement entre les indices 550 et 650.

Q4 Définir les listes `cas_moy_2020` et `cas_moy_2021` contenant le nombres de cas moyens sur chaque automne.

2. À la date de parution de l'article (décembre 2021) l'évolution du nombre de cas quotidien laissait supposer une évolution similaire à celle de l'année précédente. L'apparition de nouveaux variants plus contagieux a ensuite sensiblement modifié l'allure des courbes de taux d'incidence. La méthodologie mise en place dans l'article n'en reste pas moins intéressante !

Pour pouvoir comparer correctement les deux automnes, il faut choisir les jours "J zéro" de début de chaque vague. Voici l'extrait de l'article de presse expliquant la méthodologie des journalistes :

> *Pour définir les dates des deux vagues et le "J zéro" indiqué dans les graphiques, nous avons analysé le taux de croissance de l'incidence entre un jour J et un jour J-7. Le taux de croissance correspond à la vitesse d'augmentation de la courbe*[a]*. S'il est de 100%, l'indicateur double ; s'il est à 0%, il stagne, et s'il est négatif, l'indicateur diminue. Nous avons défini le "J zéro" de nos graphiques au moment où le taux de croissance de l'incidence dépassait les 40%*[b] *: c'est le moment où la courbe commence à augmenter davantage, correspondant ainsi à l'idée d'une vague épidémique.*

a. (vous remarquerez que ce vocabulaire n'est pas satisfaisant : c'est le nombre de cas qui augmente, pas la courbe...)

b. l'article choisissait initialement un seuil de 50%. Nous ajustons ce taux à 40% pour notre jeu de données.

Q5 Écrire une fonction `Jzero` prenant en argument une liste et renvoyant l'indice "J zéro" défini par les journalistes.

Q6 Déterminer les indices `J0_2020` et `J0_2021` de début des deux vagues automnales. Représenter ensuite comme dans l'article les deux vagues sur le même graphique depuis le jour "J-20".