BROILLET Virgile (p2103804)
GRAFF Marie (p2101141)

Formation Informatique 2022 - 2023

Statistique pour l'informatique :
Sujet n°100461

Exercice 1.A Statistique :

Nous traiterons tous les exercices de la maniéré suivante :

- *Toutes les valeurs seront arrondies à deux comme demandé*
- *Rappel de l'énoncé (en italique) pour faciliter la compréhension*
- *Réponse à la question*
- *Insertion du graphique et/ou du code si nécessaire*

Question 1)

Quel est le type de variable statistique de chacune des variables (nominale, ordinale, quantitative discrète ou continue) ?

Les index sont des variables quantitatives discrètes, les dates sont des variables ordinales, car celles-ci sont ranger par ordre croissant, et l'intégralité des mesures sont des variables quantitatives continues, car elles sont sur l'intervalle :]-∞, +∞[

Question 2)

Quel est le nombre de jours d'observation de l'échantillon ? Quel est le nombre de jours où tous les composés organiques volatils ont été mesurés ? (Indication : vous pouvez utiliser la fonction isna.)

En tous, il y a eu : 731 jours d'observations mais seulement 208, où tous les composés organiques volatils ont été mesurés.

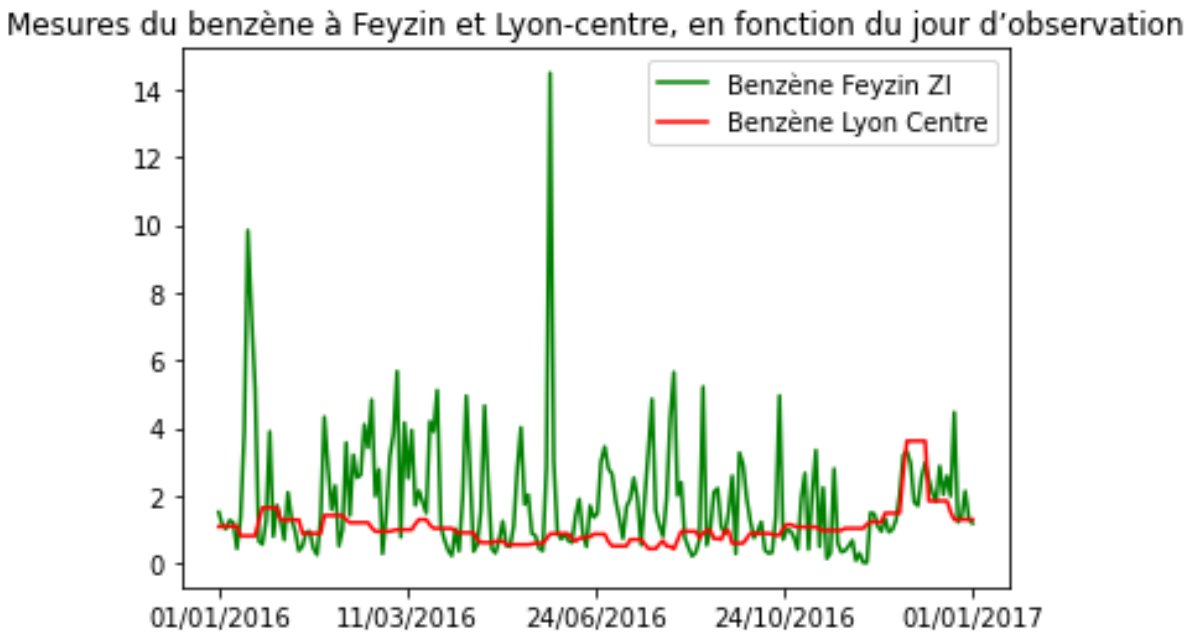
```
newAir=Air.dropna()
print(len(Air))
print(len(newAir)
```

Question 3)

Tracez sur un même graphique les mesures du benzène à Feyzin et Lyon-centre (de deux couleurs différentes), en fonction du jour d'observation. On parle parfois de série temporelle pour ce type de représentation.

Nous avons arbitrairement choisi l'affichage de ces cinq dates comprises dans le Dataframe newAir['Date'] via la fonction xtickts :

```
plt.xticks(np.linspace(0,207,5,endpoint=True)
```



Question 4)

En ignorant les jours de non observation du benzène à Feyzin (en utilisant par exemple la fonction dropna), trouvez la moyenne empirique, variance empirique, variance empirique non-biaisée et le quartile à 75 % de la mesure du benzène à Feyzin.

Moyenne empirique biaisée hors jours de non observation = 1.89
Variance empirique biaisée hors jour de non observation = 2.90
Variance empirique non-biaisée hors jour de non observation = 2.92
Quartile à 75%, hors jour de non observation = 2.59

Question 5)

Pour chaque mesure, calculer le nombre de jours de non-observation dans chacune des 105 semaines (on considèrera les derniers jours de 2015 dans la première semaine de 2016 comme des jours de non-observation).

Trouver les 4 mesures qui ont des semaines sans aucune observation (on les exclura par la suite).

Créer un DataFrame dfh avec, pour individus les semaines d'observation et pour variables : les moyennes géométriques hebdomadaires des mesures de pollution, pour chacune des 8 mesures ayant des observations pour toutes les semaines.

Les quatre mesures ayant des semaines sans aucunes observations sont :

- Benzène à Lyon Centre
- Soufre à Lyon Centre
- Soufre à Feyzin ZI
- Dioxyde de Soufre à Saint Fons

Le DataFrame 'dfh' a été créer, l'index représente la i-ème semaine, puis on trouve les valeurs des moyennes géométriques des 8 mesures complètes.

Exercice 1.B Corrélation et Régression Linéaire :

Nous traiterons tous les exercices de la maniéré suivante :

- *Toutes les valeurs seront arrondies à deux comme demandé*
- *Rappel de l'énoncé (en italique) pour faciliter la compréhension*
- *Réponse à la question*
- *Insertion du graphique et/ou du code si nécessaire*

Question 1)

Calculez les covariances non-biaisées et les corrélations des 8 relevés de dfh.

Partie Covariance :

La covariance de la M-Gé Feyzin Benzène : 0.57
La covariance de la M-Gé Feyzin Ethane : 5.34
La covariance de la M-Gé Feyzin Propane : 62.7
La covariance de la M-Gé Feyzin Propène : 6.83
La covariance de la M-Gé Vernaison Benzène : 0.35
La covariance de la M-Gé Vernaison Ethane : 4.24
La covariance de la M-Gé Vernaison Propane : 6.06
La covariance de la M-Gé Vernaison Propène : 0.68

On observe que les corrélations sont toutes égales à 1.0, car x et y valent la même chose, elles seront différentes dans les questions suivantes.

La corrélation de la M-Gé Feyzin Benzène : 1.0
La corrélation de la M-Gé Feyzin Ethane : 1.0
La corrélation de la M-Gé Feyzin Propane : 1.0
La corrélation de la M-Gé Feyzin Propène : 1.0
La corrélation de la M-Gé Vernaison Benzène : 1.0
La corrélation de la M-Gé Vernaison Ethane : 1.0
La corrélation de la M-Gé Vernaison Propane : 1.0
La corrélation de la M-Gé Vernaison Propène : 1.0

Question 2)

Effectuer un test de corrélation :

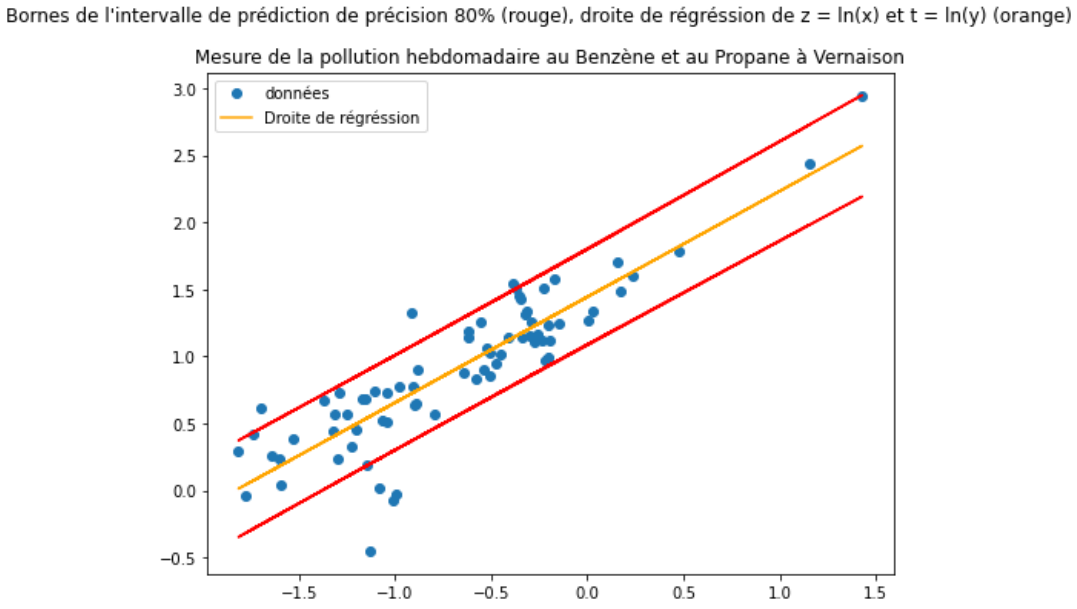
Test de corrélation entre x et y : 0.96, on en conclu que les deux mesures sont corrélées

Question 3 & 4)

Trouvez (avec la commande sm.OLS) la droite de régression de z = ln(y) en fonction de t = ln(x). Discutez la significativité statistique du résultat (en commentant les résultats de la commande fit().summary()). Tracer le nuage de points de (z, t) et la droite de régression en couleur, puis deux droites donnant les bornes de l'intervalle de prédiction au niveau 80% (d'une autre couleur).

Pour voir la droite de régression ainsi que sa prédiction, regardez le ficher .py (ligne 355 & ligne 361)

Vous trouverez le graphique ci-dessous.



Exercice 2:

ATTENTION TOUTES LES VALEURS ONT ÉTÉ GENERÉES VIA **ST.UNIFORM.RVS**

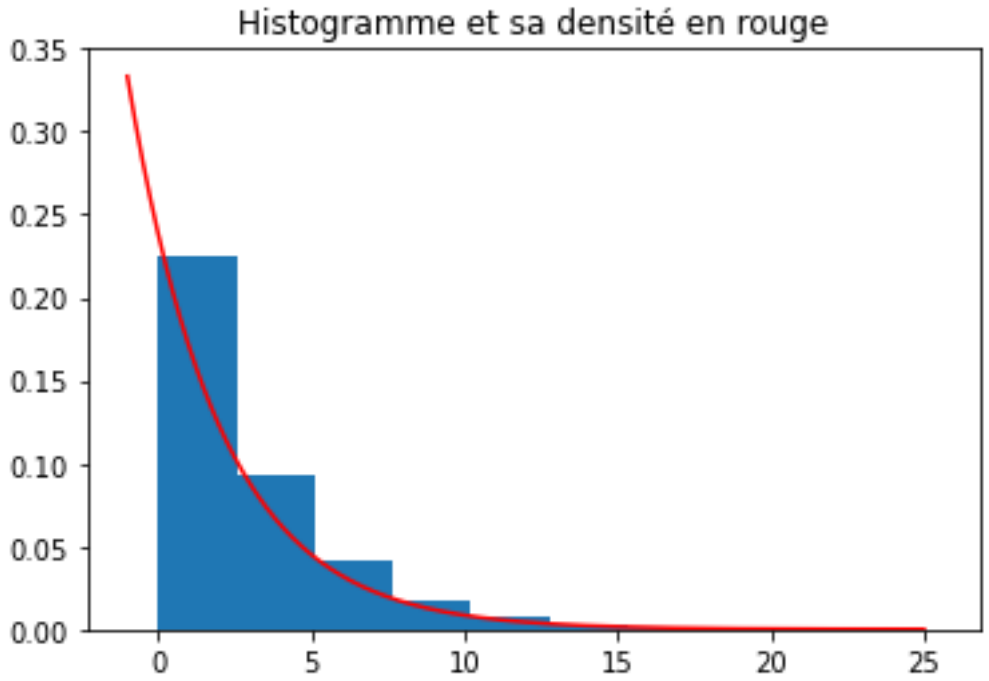
Nous traiterons tous les exercices de la maniéré suivante :

- *Toutes les valeurs seront arrondies à deux comme demandé*
- *Rappel de l'énoncé (en italique) pour faciliter la compréhension*
- *Réponse à la question*
- *Insertion du graphique et/ou du code si nécessaire*

Question 1)

En comparant un histogramme et une densité bien choisie (on tracera la courbe de la densité en rouge par-dessus l'histogramme), émettez une hypothèse sur la loi exponentielle E(λ) qui est la loi de S. Expliquer comment vous avez trouvé λ.

choix de lambda: loi exp=lamda*exp(-lamda*U)
pour simplifier la résolution de l'équation
3*log(U)=lamda*exp(-lamda*U),
-lamda/3 doit faire 1
Ainsi, on émet l'hypothèse que la loi S suit la loit E(-1/3)



Question 2)

Trouver la moyenne et la variance empirique non-biaisée de T.

Moyenne Empirique non-biaisée : 0.99
Variance Empirique non-biaisée : 1.0
Variance Empirique biaisée : 1.0

Pour l'explication du code, regardez le fichier .py (ligne 400 à 419)

Question 3)

Expliquez ce que fait le programme ci-dessus pour obtenir T. (Si nécessaire, on pourra représenter des tracés intermédiaires pour comprendre les étapes)

Le programme en question rajoute une case contenant la différence entre l'indice de case et j-1 si le le contenu de la variable x est supérieur à 3:
si x>3 est la somme de trois cases de S successives, un 2 est indiqué. si c'est le résultat de deux additions successives, un 1 est mis dans la case de y,...
Le programme permet de savoir combien de cases successives du tableau S[i] donnent une somme inférieure à trois.

Question 4)

On peut vérifier que le programme donnant T permet de simuler l'échantillon d'une loi de Poisson P(λ). Comment trouvez-vous le paramètre de la loi de T ? En comparant le diagramme en bâton de T avec le diagramme d'une loi discrète bien choisie, vérifiez graphiquement que l'hypothèse sur la loi de T est réaliste.

Un échantillon suivant la loi de poisson est un échantillon de loi de probabilité discrète, par exemple des segments. Ici, nous avons, avec le programme donnant T, un tableau de valeurs décrivant le nombre de fois où x ne subit pas une remise à zéro. Ainsi, on compare T avec la loi de Poisson de paramètre lambda=var[T].

Vous trouverez le graphique de la loi Poisson(0.9)