**Probabilités discrètes et statistiques descriptives**

**L2 Université Lyon 1 2022-2023**

# TP noté numéro 100461 à rendre avant le mercredi 21 décembre 2022 à 23h59

Il faut charger sur Tomuss une feuille de calcul python et une présentation des réponses aux questions de cette feuille dans les deux colonnes Tomuss prévues à cet effet.

**Instructions pour la synthèse de présentation des résultats du TP**

Le projet est à rédiger avec Word ou LibreOffice (ou LaTeX si vous savez déjà l'utiliser). Mais vous devrez convertir votre fichier au format **pdf**. Attention ! Quand on exporte un fichier en pdf, il manque parfois les formules mathématiques ou des graphes. Il faut soigneusement vérifier le fichier pdf et si nécessaire, essayer d'autres convertisseurs pdf (il en existe de nombreux gratuits sur le web, comme cutepdf).

Le nom des fichiers devra contenir le nom de famille des deux membres du groupe (sans accents) sous la forme Fisher_Pearson.pdf et Fisher_Pearson.py.

Rédigez soigneusement. Commentez à chaque fois vos graphes et vos résultats.

La première page devra contenir : nom, prénom, formation, année universitaire, nom de l'UE, date, numéro du sujet.

- Ajoutez des numéros de page, s'ils n'y sont pas.
- Vérifiez les graphiques : il faut qu'ils soient centrés, que leurs titres soient en français.
- Harmonisez la mise en forme des titres. Le style des paragraphes doit être justifié, pas aligné à gauche.
- Relisez pour l'orthographe (au moins deux fois), relisez pour la ponctuation.
- Écrivez le code python dans un fichier à part (colonne à part sur Tomuss). Si vous copiez du code python dans votre présentation, utilisez une police adéquate (avec un interlettrage fixe, comme `Lucida Console`).

  **Dans tous les cas, nettoyez le code des lignes inutiles, commentez-le de manière raisonnable. 3 points seront accordés à la présentation de la synthèse et à la lisibilité du code !**

Dans la suite on suppose avoir chargé les librairies suivantes :

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pan
import statsmodels.api as sm
import scipy.stats as st
```

**Exercice 1. Statistiques**

Charger le jeu de données :

```
Air=pan.read_csv('http://tinyurl.com/y39an7ef/Data100461.csv',sep='\t',na_values='-'
)
```

Ce jeu de données contient des informations issues de relevés de la pollution industrielle dans trois zones industrielles du sud lyonnais entre 2016 et 2017. À titre de comparaison, on donne aussi deux relevés à la station de Lyon-centre. Ils sont mesurés en microgrammes par mètre cube d'air. Il contient 12 relevés de 5 polluants :

a. le dioxyde de soufre est un polluant industriel commun dont la mesure sert à déclencher les alertes en cas de pollution industrielle ; il est mesuré à Feyzin Z.I., Saint-Fons et Lyon-centre. Attention, à cause des incertitudes de mesures de la méthode utilisée, ces mesures peuvent être négatives et l'observatoire de la pollution préfère garder ces valeurs pour ne pas fausser la moyenne de pollution en ne retenant que les erreurs positives.

b. un composé organique volatil cancérigène : le benzène ; il est mesuré à Feyzin Z.I., Lyon-centre et Vernaison Z.I. ;

c. d'autres composés organiques volatils : l'éthane, le propane et le propène ; ils sont mesurés à Feyzin Z.I. et à Vernaison Z.I.

Les valeurs indiquées "-" dans le fichier, ou NA ou nan après importation par `pandas`, sont les valeurs non disponibles, correspondant à des jours de non-observation.

**On donnera les réponses numériques arrondies avec DEUX décimales**.

**A. Statistiques Descriptives et Informations générales**

1. Quel est le type de variable statistique de chacune des variables (nominale, ordinale, quantitative discrète ou continue) ?
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`.)
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.
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.
5. Comme les mesures sont positives, il est pertinent de considérer parfois des moyennes géométriques $\sqrt[n]{x_1 \cdots x_n}$ (son logarithme est la moyenne empirique usuelle des logarithmes). Dans la suite, on va s'intéresser aux moyennes géométriques hebdomadaires pour limiter l'impact des jours de non observation. On considère des semaines allant du lundi au dimanche suivant et on remarque que le premier lundi de l'échantillon est le 4 janvier 2016. 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.

**Dans la suite, on travaille sur les données hebdomadaires du data.frame `dfh`, créé à la dernière question.**

## B. Corrélation et Régression Linéaire

1. Calculez les covariances non-biaisées et les corrélations des 8 relevés de `dfh`.
2. Soit $x$ l'échantillon des mesures de la pollution hebdomadaire au Benzène à Vernaison et $y$ l'échantillon des mesures de la pollution hebdomadaire au Propane à Vernaison. On cherche à savoir si elles sont corrélées. Quelle hypothèse doit-on faire pour effectuer un test de Pearson ? Effectuez un test de corrélation. Que concluez-vous ?
3. 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).
4. Tracer les nuages de points des mesures Benzène $x$ et en Propane $y$ à Vernaizon avec la courbe de régression (en supposant qu'il suffit de prendre l'image par exp de la régression linéaire de $t, z$, ce qui n'est qu'une première approximation) et les bornes des intervalles de prédiction au niveau 80% pour ces mesures.

   En utilisant cette approximation, calculer un intervalle de prédiction pour la mesure de Propane si la mesure en Benzène est au seuil réglementaire d=2. Est-ce que cela vous semble justifier que le Benzène serve de référence pour la mesure de pollution des autres hydrocarbures comme le Propane ?

**Exercice 2.**

On rappelle que la fonction `st.uniform.rvs` permet de simuler une loi uniforme continue en obtenant un vecteur dont les nombres sont uniformément répartis dans l'intervalle spécifié.

1. On donne les commandes suivantes :

```
U=st.uniform.rvs(0,1,size=10000);
S=(-np.log(U)*3);
```

   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 $\mathcal{E}(\lambda)$ qui est la loi de $S$. Expliquer comment vous avez trouvé $\lambda$.

2. On considère le programme suivant :

```
x=0; y= [0]; j= 1
for i in range(len(S)):
    x= x + S[i]
    if(x>3):
        x = 0
        y = np.concatenate((y,[i+1-j]))
        j = i+2
T = y[1:len(y)]
```

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

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

4. On peut vérifier que le programme donnant $T$ permet de simuler l'échantillon d'une loi de Poisson $\mathcal{P}(\lambda)$. 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.

5. Pour effectuer un test du $\chi^2$, on veut se ramener à une loi sur $[\![0, 5]\!]$. Construire un vecteur correspondant à la variable $R$ dont les composantes sont $R_i = \min(T_i, 5)$. Soit $X$ de loi de Poisson $\mathcal{P}(\lambda)$ avec $\lambda$ trouvé à la question précédente, construire un vecteur $RTh$ de longueur 6 contenant les probabilités $RTh_i = P(min(X, R) = i)$ pour $i = 0, \cdots, 5$.

   Tracer sur un même graphique les diagrammes à barres de $T$ et $R$ et des croix rouges de coordonnées $(i, RTh_i)$ pour $i = 0, \cdots, 5$.

6. Pour tester votre hypothèse du 4. sur la loi de $T$, on s'est donc ramené à une variable à support fini $R$ : Tester donc, en utilisant un test du $\chi^2$ d'adéquation à une loi discrète si la variable $R$ suit la loi $RTh$. (On calculera en particulier la $p$-valeur du test pour discuter de la significativité statistique du résultat).