**Université de Montpellier** **HLMA408: Traitement de données**
**2020/2021** **Joseph Salmon**

---

**TP n° 3 : (TP Noté) Moindres carrés**

---

Objectifs du TP : Test d'égalité de moyennes, permutations aléatoires et hypothèse nulle, modèles linéaires.

Consignes :

Pour ce travail vous devez déposer un **unique** fichier sur le moodle du cours HLMA408.

Attention, vous devrez veiller à ce que le format soit correct : si le fichier ne peut pas être ouvert par le correcteur avec `jupyter-notebook`, la note sera de zéro : **cette vérification vous incombe** !

Pour faciliter la correction, le nom du fichier devra respecter le format suivant :

`nom_fichier = tp_note3_hlma408_gr_nbgr_prenom_nom.ipynb`, en minuscule, sans accent ni espace. Vous remplirez votre nom, prénom, le numéro de groupe qui vous concerne (remplacer `nbgr` par C, D ou E) de manière adéquate [1].

**Un point de malus sera appliqué pour les fichiers dont le format est erroné**.

Vous devez charger votre fichier sur Moodle, avant le **vendredi 16/04/2021, 23h59**. La note totale est sur **20** points, répartis comme suit :

- qualité des réponses aux questions : **14** pts,
- qualité de rédaction et d'orthographe : **1** pt,
- qualité des graphiques (légendes, couleurs) : **2** pts
- qualité d'écriture du code (noms de variable clairs, commentaires, code synthétique, etc.) : **1** pt
- Rendu reproductible et absence de bug : le code doit s'exécuter sur la machine du correcteur sans manipulation de sa part (par exemple le correcteur n'est pas supposé aller chercher les fichiers sur internet, les enregistrer, etc.). On veillera donc à ce que les traitements de donnée soient automatisés et ne requièrent pas d'interversion supplémentaire **2** pts.

Les personnes qui n'auront pas soumis leur devoir sur Moodle avant la limite obtiendront **zéro**.

### Exercice 1. Préliminaires (0.5pt)

1) (**0.25pt**) Construire la chaîne de caractères `nom_fichier` données ci-dessus, et donner sa taille.
2) (**0.25pt**) Calculer $\alpha$ qui vaut la longueur de la chaîne de caractère `mon_fichier` divisé par 1000 (par exemple pour `nom_prenom = Salmon_Joseph`, on trouve $\alpha = 0.041$, soit 4.1 %) et donner le quantile $(1 - \alpha)$ d'une loi gaussienne centrée réduite.

### Exercice 2. Dés et tirages aléatoires (2pts)

On se propose de simuler numériquement le fonctionnement d'un dé à 4 faces (numérotées de 1 à 4), et de vérifier statistiquement que la simulation est satisfaisante.

1) (**0.75pt**) Créer une fonction appelée `echantillon_de4` en `Python` qui prend en entrée `n_samples`, et qui renvoie en sortie un vecteur (`numpy array`) de taille `n_samples`, dont les valeurs sont tirées au hasard et uniformément entre 1 et 4.
2) (**0.25pt**) Tirer un échantillon de taille `n_samples` = 1000 avec cette fonction, et donner les fréquences empiriques des 4 faces.
3) (**1pt**) Proposer une méthodologie pour valider par un test statistique si votre fonction est bien valide, et conclure sur la qualité de votre méthode.

### Exercice 3. Prairies et rendement agricole (1pt)

Dans cet exercice, on veut faire une analyse de la variance pour vérifier l'influence du type de sol sur le rendement fourrager. On dispose de 30 observations de parcelles de prairie pour lesquelles on a mesuré la variable `rendement` (en tonnes) et on donne la variable `parcelle` qui indique le type de sol (codé par 1, 2 ou 3).

1. Exemple : Joseph Salmon est dans le groupe C, son TP s'appelle `tp_note3_hlma408_gr_C_joseph_salmon.ipynb`

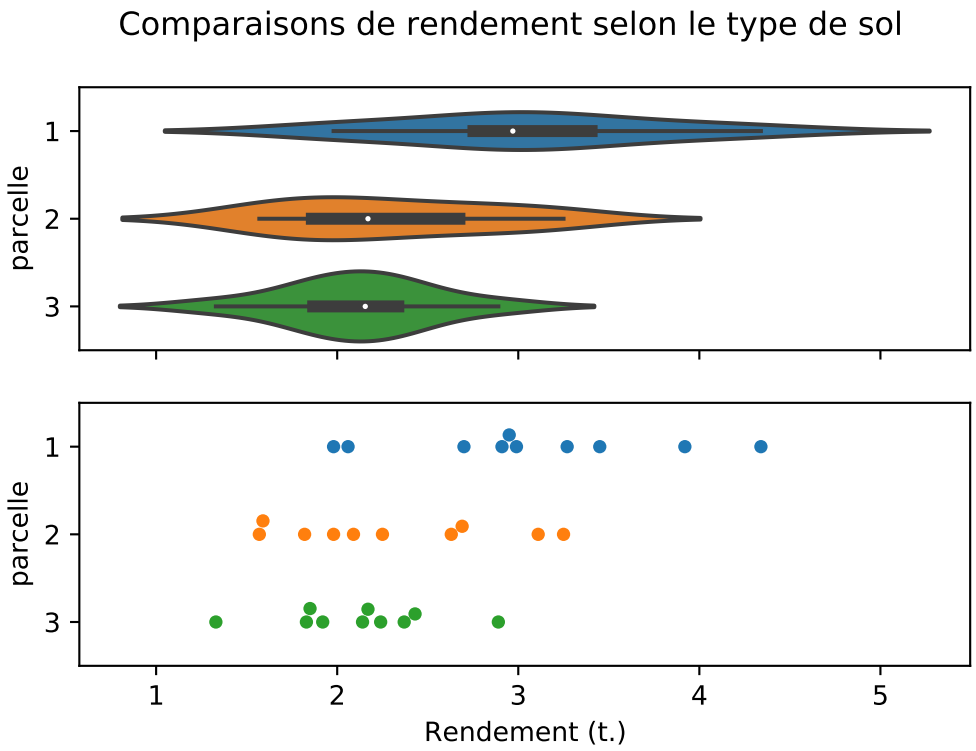


FIGURE 1 – *swarm* (🇫🇷 : *essaim*) et *violin* (🇫🇷 : *violon*)

1) Importez le jeu de données `prairie.txt` [2] dans une table nommée `prairie`.
2) (**1pt**) Reproduire avec Python le graphique de la Figure 1
   Tracer avec `subplots` un graphique qui contient deux sous-graphiques du rendement des parcelles en fonction du type de sol (encodé par `parcelle`) : en haut, un diagramme en violon ; en bas, un graphique `swarmplot` des données [3].

**EXERCICE 4. Impact d'un traitement sur la croissance des plantes (2pts)** Une expérience de comparaison de deux traitements de plantes repose sur l'étude d'un échantillon de 20 plantes : 10 plantes ont été sélectionnées au hasard pour subir le traitement 1 et les 10 plantes restantes ont subi le traitement 2. On va étudier la masse (en grammes) des plantes après traitement.

1) Créez deux vecteurs `echA` et `echB` qui contiennent respectivement les données des poids avec le traitement A et le traitement B :

```
14.4, 14.7, 13.2, 12.1, 18.7, 15.0, 13.3, 17.8, 16.6, 15.0 (traitement A)
25.6, 17.7, 19.0, 26.7, 22.6, 19.1, 22.9, 21.0, 25.7, 23.7 (traitement B)
```

2) (**1pt**) Donner un intervalle de confiance bilatéral de la différence des espérances entre les deux traitements (on prendre un niveau de confiance $1-\alpha$ avec le même $\alpha$ qu'à l'EXERCICE 1). Comparer les résultats obtenus avec et sans l'hypothèse de variances égales.
3) (**1pt**) Tester l'égalité des moyennes des deux groupes de plantes qui reçoivent le traitement A et B, Optez pour un test bilatéral et comparez les résultats obtenus avec et sans l'hypothèse de variances égales. Donnez $\mathcal{H}_0$, $\mathcal{H}_1$ et la $p$-valeur de ces tests. Conclure.

**EXERCICE 5. Hospitalisation : Répartition des entrées (2pts)**
Dans cet exercice on souhaite savoir si les entrées à l'hôpital pour une certaine maladie (la maladie **A**) sont réparties au hasard dans l'année ou bien si certains mois sont plus propices à la maladie **A**. On

2. http://josephsalmon.eu/enseignement/datasets/prairie.txt
3. cf. https://seaborn.pydata.org/generated/seaborn.swarmplot.html

examine le mois d'entrée d'un échantillon de 120 porteurs de la maladie **A**. Les résultats sont contenus dans le fichier `Hospit.csv` [4].

Écrire un code qui vous permettra de répondre à la question suivante : Peut-on affirmer avec un risque $\alpha$ (le même $\alpha$ qu'à l'EXERCICE 1) que "les entrées pour la maladie **A** ne se font pas au hasard dans l'année" ? Vous inclurez des commentaires dans votre code et vous indiquerez clairement $\mathcal{H}_0$, $\mathcal{H}_1$, le test utilisé et la $p$-value de ce test ainsi que la conclusion que vous pouvez en tirer.

**Exercice 6. Arbres : taille et volume** (**6.5pts**)

Un étudiant en techniques forestières veut utiliser la régression linéaire pour estimer le volume en bois utilisable d'un arbre debout en fonction de l'aire du tronc mesurée à 25 $cm$ du sol. Il a choisi au hasard 10 arbres et a mesuré, à la base, l'aire correspondante (en $cm^2$). Il a par la suite enregistré, une fois l'arbre coupé, le volume correspondant en $m^3$.

Le fichier `arbres.txt` contient les données. Les variables sont `vol` et `aire` qui représentent respectivement le volume utilisable et l'aire à la base du tronc.

Le but de cet exercice est d'étudier la variable `vol` en fonction de la variable `aire`.

1) (**0.25pt**) Importez automatiquement le jeu de données `arbres.txt` [5] dans un dataframe que vous nommerez `df_arbres`.
2) (**1pt**) Ajustez le modèle linéaire qui explique la variable `vol` (en ordonnée) par la variable `aire` (en abscisse). Donner la valeur de la pente estimée ainsi que celle de l'ordonnée à l'origine. [6]
3) (**1pt**) Représentez le nuage des points du volume (en ordonnée) en fonction de l'aire (en abscisse) ainsi que la droite d'ajustement des moindres carrés. On prendra soin aux légendes, au titre, aux noms des axes, etc.
4) (**0.5pt**) Donnez la proportion de la variance expliquée par le modèle linéaire.
5) (**0.75pt**) Afficher une courbe représentant un estimateur à noyau des résidus (après les avoir centrés et réduits). On veillera à ce que l'aire sous la courbe vaille 1.
6) (**1pt**) Proposez un intervalle de prédiction pour le volume correspondant à une aire de $465cm^2$ pour un niveau de confiance de $1-\alpha$ (avec $\alpha$ la valeur obtenu à l'EXERCICE 1).
7) (**2pt**) Afficher sur un graphique les données brutes, la droite de régression obtenue par les moindres carrés, la prédiction donnée par le modèle pour la valeur $465cm^2$. Enfin, proposer une représentation graphique de l'intervalle de confiance pour le niveau de confiance $\alpha$ calculé à la question précédente.

**Exercice 7. Pollution en Occitanie** (**2pts Bonus**) Reprendre la base de données sur les principaux polluants en Occitanie [7].

Proposer une comparaison approfondie du niveau de pollution entre Toulouse et Montpellier sur le niveau de NO2. On pourra s'intéresser au niveau de pollution pour certaines saisons seulement, pour certains jours de la semaine/week-end, pour la journée / pour la nuit, etc. Ces comparaisons seront appuyées d'un argumentaire graphique et statistique (test/intervalle de confiance) de votre choix.

4. http://josephsalmon.eu/enseignement/datasets/Hospit.csv
5. http://josephsalmon.eu/enseignement/datasets/arbres.txt
6. On pourra utiliser le package `statsmodel`, *cf.* http://www.statsmodels.org/stable/regression.html ou tout autre fonction
7. http://josephsalmon.eu/enseignement/datasets/Mesure_journaliere_Region_Occitanie_Polluants_Principaux.csv