BUT1 – S.A.E. S2-02 ET S2-04
STATISTIQUES ET MÉTHODES NUMÉRIQUES

IUT DE NANTES – DÉPARTEMENT D'INFORMATIQUE

Contact : François Simonneau, Email : francois.simonneau@univ-nantes.fr

L'évaluation se découpe en deux parties : une première partie relative à l'exploitation de ce qui a été vu sur la ressource « Outils numériques pour les statistiques descriptives » et une deuxième partie relative à l'exploitation de ce qui a été fait sur python sur la ressource « Méthodes numériques » :

L'évaluation sera individualisée et elle s'effectuera par le biais d'une évaluation sur python reprenant l'ensemble de ce sujet. Chaque question du sujet sera traitée par l'implémentation d'une fonction propre.

Partie 1 – Statistiques

Vous partirez des données suivantes :

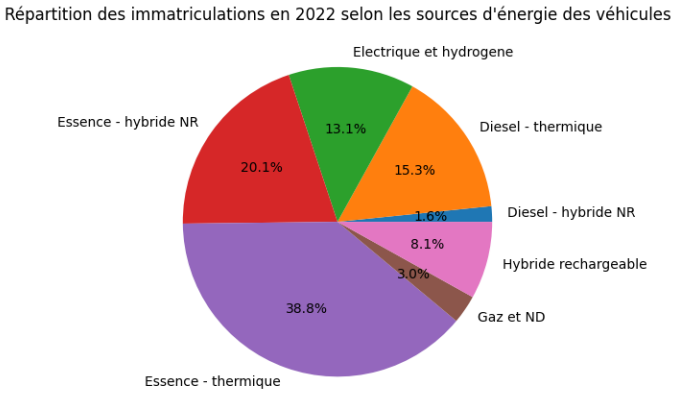
- `RGC_2013.csv` donnant des informations sur l'ensemble des communes de France
- `data_immat_traitees.csv` donnant des informations sur les immatriculations de véhicules neufs entre 2010 et 2022. Il y a en fait dix fichiers à exploiter pour en extraire des dataframes à concaténer (l'opération effectuant cette concaténation est réalisée dans le fichier déposé sur madoc avec l'exploitation d'un dictionnaire de dataframes)
- `data_bornes_traitees.csv` reprenant des données utilisées sur la ressource « Exploitation d'une base de données » avec un traitement sommaire de celles-ci où les lignes ne faisant pas apparaître de `code_insee_commune` ont été corrigées ou majoritairement supprimées et où un identifiant de département `dep_b` a été ajouté.

Vous devrez ici être en mesure d'effectuer l'ensemble des questions figurant dans cette partie.

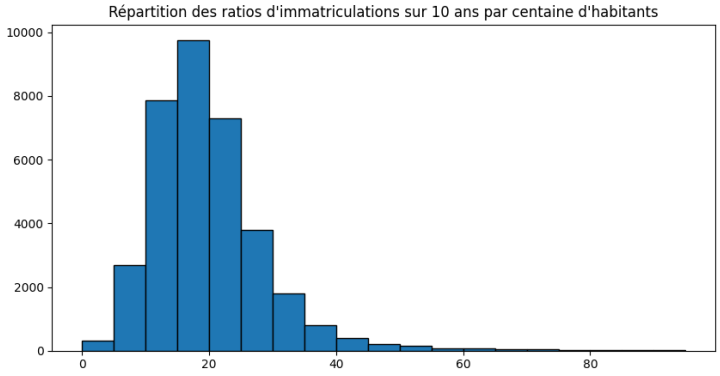
Q. 1.1. Implémenter la fonction `calcul_ratio(list_dep)` retournant une `serie` indexée par les chaînes de caractère désignant les numéros de départements et donnant pour une liste de départements passée en argument (sélectionnés parmi les départements des Pays de la Loire) le ratio présentant le nombre de voitures exploitant l'energie `Electrique et hydrogene` par point de charge sur chaque département. Le nombre de voitures est calculé en effectuant la somme de toutes les immatriculations neuves sur l'ensemble de la période (approche discutable mais en l'état des données disponibles on va dire que c'est pas si mal). Vous devez pouvoir obtenir le tableau présenté ci-dessous en prenant en entrée l'ensemble des départements des Pays de la Loire.

dptmt	Ratio
44	4.770395
49	2.664015
53	5.982005
72	5.063337
85	5.014318

Q. 1.2. Implémenter la fonction `repart_energie(annee)` renvoyant une `serie` indexée par les chaînes de caractère donnant les différentes sources d'énergie et donnant sur une année passée en paramètre le nombre total sur l'ensemble de la France des immatriculations selon le type d'énergie utilisé par le véhicule. Ce n'est pas demandé ici mais si on affichait le graphique montrant la répartition correspondante en 2022 cela donnerait le graphique ci-dessous.



Q. 1.3. Calculer pour chaque commune le ratio du nombre d'immatriculation (tous types d'énergie confondus) par centaine d'habitants et sur dix ans (on effectuera le calcul du ration sur l'ensemble de la période et on utilisera un prorata pour le ramener à un ratio attendu sur 10 ans). On aura pris soin préalablement de retirer de l'étude les communes dont la population est renseignée comme étant égale à 0. On restreint l'étude qui suit aux communes ayant un ratio inférieur ou égal à 100 (pour éviter quelques cas particuliers comme celui de la petite commune d'Avrigny où la présence d'une grande plateforme logistique automobile perturbe les données). La série des ratios obtenus sur toutes les communes constituera le premier élément renvoyé par la fonction `repart_ratio()` (pour indication, même si ce n'est pas demandé on donne ci-dessous la distribution attendue pour cette série) Calculer ensuite sur les communes restantes le ratio sur l'ensemble des communes regroupées par statut administratif (attention ce n'est pas une moyenne des ratios précédents qu'il convient d'effectuer et on effectuera toujours le calcul du ratio sur l'ensemble de la période en utilisant un prorata pour le ramener à un ratio attendu sur 10 ans). La série indexée par le statut administratif sera le deuxième élément renvoyé par la fonction `repart_ratio()`.



On souhaite effectuer une prévision de l'évolution du taux représenté par les véhicules à énergie électrique et hydrogène parmi l'ensemble des immatriculations de véhicules neufs jusqu'en 2050 (où on établira une hypothèse d'un taux de saturation du marché quant à ce taux : par exemple 80%)

Q. 1.4. Implémenter une fonction `get_tx_elec()` renvoyant la `serie` des taux annuels qu'ont représenté les véhicules à énergie électrique et hydrogène parmi l'ensemble des immatriculations de véhicules neufs entre 2010 et 2022. La série devra être seulement indexée par les années.

Dans la question suivante on va utiliser le vecteur de valeurs `x` correspondant aux années (`np.arange(2010,2023)`) et le vecteur `y` correspondant à une série du type de celle renvoyée par la fonction de la question précédente (avec la période qui peut-être raccourcie) et le vecteur de valeur `x` correspondant aux années placées en index de cette même série. On veut mettre en place un modèle à partir d'une méthode de régression exploitant une fonction logistique du type :

$$f(x) = \frac{tx\_sat}{1 + e^{ax+b}}$$

Pour cela on va exploiter notre savoir-faire sur le modèle de régression linéaire en remarquant qu'un changement de variable nous permet de passer d'une liaison attendue via la fonction logistique à une liaison linéaire :

$$y = \frac{tx\_sat}{1 + e^{ax+b}} \Leftrightarrow ln(\frac{tx\_sat}{y} - 1) = ax + b$$

La méthode de régression linéaire exploitée avec le vecteur `x` et le vecteur correspondant aux valeurs de $ln(\frac{tx\_sat}{y} - 1)$ permettra donc d'obtenir les valeurs de a et b nécessaires pour la modélisation avec la fonction logistique.

Q. 1.5. Implémenter une fonction `mod_log_evol_tx_elec(tx_sat,x_mod_log,serie_tx_elec)` renvoyant en premier élément un vecteur de taux estimés par le modèle logistique pour des dates `x_mod_log` passées en argument et en deuxième élément le coefficient a dans le modèle retenu.

Partie 2 – Méthodes numériques

On souhaite dans un premier temps modéliser l'évolution de la part que représentent les véhicules à énergie électrique et hydrogène en exhibant à partir du modèle logistique retenu une relation de récurrence permettant de définir le taux sur une année à partir du taux de l'année précédente d'après ce modèle. Pour cela on part des relations suivantes :

$$f(x) = \frac{tx\_sat}{1+e^{ax+b}} \text{ et } f(x+1) = \frac{tx\_sat}{1+e^{ax+b+a}}$$

Puis on étudie leurs inverses :

$$\frac{1}{f(x)} = \frac{1+e^{ax+b}}{tx\_sat} \text{ et } \frac{1}{f(x+1)} = \frac{1+e^{ax+b+a}}{tx\_sat}$$

Ou encore :

$$\frac{e^a}{f(x)} = \frac{e^a+e^{ax+b+a}}{tx\_sat} \text{ et } \frac{1}{f(x+1)} = \frac{1+e^{ax+b+a}}{tx\_sat}$$

Et on étudie alors la différence :

$$\frac{1}{f(x+1)} - \frac{e^a}{f(x)} = \frac{1-e^a}{tx\_sat}$$

En considérant un taux donné u_0 pour une année de départ on souhaite dès lors pouvoir définir une suite (u_n) (où n parcourra un ensemble d'années entières suivantes) par récurrence à partir de la relation suivante :

$$\frac{1}{u_{n+1}} - \frac{e^a}{u_n} = \frac{1-e^a}{tx\_sat}$$

Q. 2.1. Conclure les calculs précédents pour déterminer la relation de récurrence permettant d'obtenir u_{n+1} en fonction de u_n. Exploiter la relation obtenue en implémentant une fonction `mod_log_rec(u0,n,tx_sat,a)` renvoyant la liste composée des `n+1` premières valeurs de la suite

On souhaite maintenant modéliser l'évolution du taux représenté par les véhicules à énergie électrique et hydrogène parmi l'ensemble des immatriculations de véhicules neufs jusqu'en 2050 par un polynôme interpolant la fonction logistique. Pour cela nous allons utiliser la méthode d'interpolation de Lagrange avec les points d'interpolation de Tchebychev avec N points définis ici pour un intervalle $[a,b]$:

$$x_k = \frac{a+b}{2} + \frac{b-a}{2} \times cos\left(\frac{2k-1}{2N}\pi\right), \quad k = 1, \cdots, N$$

Q. 2.2. Implémenter une fonction `BaseLagrange(x,listX,i)` qui prend en entrée une valeur `x`, une liste de nombres `listX` (deux à deux différents mais qui ne correspondent pas nécessairements aux points de Tchebychev) et un entier positif `i` strictement plus petit que la longueur de `listX` et qui renvoie l'image du vecteur `x` par la fonction polynomiale

$$\begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & \prod_{\substack{j=1 \\ j\neq i}}^{n} \frac{x-x_j}{x_i-x_j} \end{array}.$$

où les x_i sont les éléments de `listX`.

Q. 2.3. Implémenter une fonction `InterLagrange(x,listX,listY)` qui prend en entrée une valeur `x` et deux listes de nombres `listX` et `listY` de même taille n et qui renvoie l'image de

x par le polynôme interpolateur de Lagrange de degré au plus n passant les points (x_i, y_i) pour x_i = `listX`$[i]$ et y_i = `listY`$[i]$.

On souhaite pouvoir dire à quelle date le taux sera égal à une certaine valeur attendue en s'appuyant sur la fonction précédente et en utilisant deux méthodes de résolution.

Q. 2.4. Implémenter une fonction mettant en oeuvre la méthode de dichotomie(`dicho(a,b,f,e)`) en renvoyant la liste des solutions approchées obtenues successivement avec comme dernier élément de la liste la première solution obtenue $\bar{x}$ telle que $f(\bar{x}) < e$ (recherche initialisée en considérant que la valeur cherchée se situe entre `a` et `b` sans que ces deux valeurs ne soient intégrées à la liste des résultats)

Q. 2.5. Implémenter une fonction mettant en oeuvre la méthode de la fausse position (`fausse_pos(a,b,f,e)`) en renvoyant la liste des solutions approchées obtenues successivement avec comme dernier élément de la liste la première solution obtenue $\bar{x}$ telle que $f(\bar{x}) < e$ (recherche initialisée en considérant que la valeur cherchée se situe entre `a` et `b` sans que ces deux valeurs ne soient intégrées à la liste des résultats)

Remarque : Même si ceci ne sera pas évalué vous pourrez vous assurer de la cohérence des résultats obtenus dans cette partie en confrontant les résultats obtenus dans l'estimation de la date à laquelle on atteindrait un taux de 50% :

— Avec recherche de la première année où la valeur dépasse ce taux dans la liste renvoyée par `mod_log_rec(u0,n,k,a)`

— Avec dernières valeurs des listes renvoyées par `dicho(a,b,f,e)` et `fausse_pos(a,b,f,e)` en prenant pour fonction `f` la fonction `lambda x:InterLagrange(x,listX,listY)-0.5` où `listX` correspond à une liste de points d'interpolation de Tchebychev et `listY` correspondant à la liste d'images de ces points par la fonction `mod_log_evol_tx_elec`.