BUT1 – S.A.E. S2-04
STATISTIQUES

IUT DE NANTES – DÉPARTEMENT D'INFORMATIQUE

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

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 – Independance

1.1. **Cas théorique**. On imagine ici une entreprise ayant trois sites de production (appelés « Site 1 », « Site 2 » et « Site 3 ») et qui mène une étude sur les causes des arrêts sur ses chaînes de production. Quatre origines de défaillances ont été identifiées menant à ces arrêts et l'entreprise souhaite ici savoir s'il y a des tendances particulières qui apparaissent sur l'origine d'une défaillance selon le site de production. Autrement dit la problématique est de savoir si la défaillance a la même chance d'avoir telle ou telle origine quel que soit le site de production (indépendance des deux variables) et dans le cas contraire identifier les cas les plus singuliers (sur-représentation ou sous-représentation d'une origine dans un site par rapport à ce qui est observé sur l'ensemble des sites.

Pour mener cette étude des données ont été collectées sur l'échantillon des 250 derniers arrêts constatés sur l'ensemble des trois sites.

Effectifs observés	Electrique	Humaine	Mécanique	Réseau	Total
Site 1	10	10	14	22	56
Site 2	28	28	31	22	109
Site 3	25	9	30	21	85
Total	63	47	75	65	250

Pour mener cette étude une première idée naturelle peut être d'observer la répartition des origines des défaillances sur chacun des sites et de pouvoir comparer chacune à ce qui a été observé au niveau de l'ensemble de l'entreprise.

Répartitions observées	Electrique	Humaine	Mécanique	Réseau	Total
Site 1	17.9%	17.9%	25%	39.3%	100%
Site 2	25.7%	25.7%	28.4%	20.2%	100%
Site 3	29.4%	10.6%	35.3%	24.7%	100%
Total	25.2%	18.8%	30%	26%	100%

On souhaiterait établir un tableau avec les effectifs qu'on aurait théoriquement dû observer sur chaque croisement pour pouvoir affirmer « à 100% » que la fréquence de telle ou telle origine est

indépendante du site de production. Ce serait le cas si les fréquences observées dans le tableau étaient identiques sur chaque ligne et en particulier identique à celle observée sur l'ensemble de la « population » soit la ligne représentée par le total sur l'ensemble des trois sites de production (ceci va donc constituer une norme pour le calcul de ces effectifs théoriques.

Ainsi par exemple l'effectif $n^t_{(S1,E)}$ qu'on aurait théoriquement dû observer, sous hypothèse d'indépendance, sur le croisement entre les modalités « Site 1 »et « Electrique » doit respecter les égalités suivantes :

$$\frac{n^t_{(S1,E)}}{Tot_{(S1)}} = 25.2\% \Leftrightarrow \frac{n^t_{(S1,E)}}{Tot_{(S1,E)}} = \frac{Tot_E}{Tot_{echantillon}} \Leftrightarrow n^t_{(S1,E)} = \frac{Tot_{S1} \times Tot_E}{Tot_{echantillon}}$$

Ce raisonnement se généralise pour obtenir l'effectif théorique de chaque croisement en se basant sur les totaux observés pour chaque modalité en ligne et chaue modalité en colonne et l'effectif total de l'échantillon :

$$n^t_{(mod_{ligne}, mod_{colonne})} = \frac{Tot_{ligne} \times Tot_{colonne}}{Tot_{echantillon}}$$

Effectifs théoriques	Electrique	Humaine	Mécanique	Réseau
Site 1	14.1	10.5	16.8	14.6
Site 2	27.5	20.5	32.7	28.3
Site 3	21.4	16	25.5	22.1

Pour étudier la façon dont l'effectif observé sur chaque croisement s'est plus ou moins éloigné de l'effectif théorique attendu sous hypothèse d'indépendance, la première idée est de calculer des écarts entre ces deux effectifs pour chaque croisement. Néanmoins de manière générale la première attente est de prendre une décision sur le fait qu'on considère ou non que les deux variables sont globalement indépendantes ou non (même si dans le cadre de cet exercice on ne pourra pas apporter une réponse à cette question). On voudrait donc pouvoir cumuler tous ces écarts (en faire la somme) de sorte à rendre compte d'un écart global entre la situation observée et la situation théorique. Or si on effectue de simples écarts ceux-ci vont être tantôt négatifs, tantôt positifs et leur somme sera nulle. Pour éviter ceci on va élever tous ces écarts au carré. Enfin il convient de relativiser les écarts selon l'effectif théorique en divisant par celui-ci (on considère qu'un écart entre un effectif observé de 10 et un effectif théorique de 12 est davantage le témoin d'une dépendance qu'un même écart entre un effectif observé de 100 et un effectif théorique de 102). Le calcul final de ce qu'on appelle des contributions (chaque croisement contribue plus ou moins au fait de rejeter l'hypothèse d'indépendance) s'établit donc ainsi :

$$Contrib_{(mod_{ligne}, mod_{colonne})} = \frac{\left(n^{obs} - n^t\right)^2}{n^t}$$

Contributions	Electrique	Humaine	Mécanique	Réseau
Site 1	1.2	0.03	0.47	3.8
Site 2	0.01	2.75	0.09	1.42
Site 3	0.6	3.05	0.79	0.05

1.2. **Cas pratique SAE.**

Q. 1.1. Implémenter la fonction `calcul_eff_obs()` renvoyant un Dataframe avec les résultats attendus au format suivant :

Style ...	Radio 1	Radio 2	Radio 3	Radio 4	Radio 5	Radio 6
Electro	12	11	39	42	18	19
Hip-Hop ...	4	6	46	18	31	58
Indie	4	9	14	8	15	7
Jazz	17	15	19	9	14	9
Musique ...	23	16	11	3	2	1
Pop	5	27	31	17	37	67
Rock	20	34	33	15	33	30
Variété	6	54	28	16	20	27

Q. 1.2. Implémenter la fonction `calcul_eff_theo()` renvoyant un Dataframe avec les résultats attendus au format suivant :

index	Radio 1	Radio 2	Radio 3	Radio 4	Radio 5	Radio 6
Electro	12.831	24.252	31.161	18.048	23.97	30.738
Hip-Hop ...	14.833	28.036	36.023	20.864	27.71	35.534
Indie	5.187	9.804	12.597	7.296	9.69	12.426
Jazz	7.553	14.276	18.343	10.624	14.11	18.094
Musique ...	5.096	9.632	12.376	7.168	9.52	12.208
Pop	16.744	31.648	40.664	23.552	31.28	40.112
Rock	15.015	28.38	36.465	21.12	28.05	35.97
Variété	13.741	25.972	33.371	19.328	25.67	32.918

Q. 1.3. Implémenter la fonction `calcul_contrib()` renvoyant un Dataframe avec les résultats attendus au format suivant :

index	Radio 1	Radio 2	Radio 3	Radio 4	Radio 5	Radio 6
Electro	0.053819...	7.241279...	1.972013...	31.78736...	1.486896...	4.482420...
Hip-Hop ...	7.911675...	17.32006...	2.763249...	0.393141...	0.390620...	14.20389...
Indie	0.271634...	0.065933...	0.156260...	0.067929...	2.909814...	2.3693446
Jazz	11.81594...	0.036717...	0.023532...	0.248246...	0.000857...	4.570622...
Musique ...	62.90290...	4.210073...	0.152987...	2.423580...	5.940168...	10.28991...
Pop	8.237072...	0.682630...	2.296697...	1.822720...	1.045984...	18.02364...
Rock	1.655026...	1.112910...	0.329253...	1.773409...	0.873529...	0.990850...
Variété	4.360896...	30.24675...	0.864452...	0.573033...	1.252391...	1.063938...

Q. 1.4. Implémenter une fonction `analyse_contrib(n)` renvoyant une liste de `n` tuples informant sur les plus grandes contributions au format suivant

`(mod_style,mod_radio,signe,contrib)` où `signe` sera représenté par le caractère `'+'` si l'effectif observé est supérieur à l'effectif théorique (tendance à une certaine attraction entre ces deux modalités) et par le caractère `'-'` si l'effectif observé est inférieur à l'effectif théorique (tendance à une certaine répulsion entre ces deux modalités).

Partie 2 – Prevision

Les fonctions à implémenter en fin de sujet appliqueront les méthodes décrites dans cette présentation mais sur des données issues de Médiamétrie, donnant l'audience cumulée pour l'ensemble des radios au niveau national, soit la part des personnes interrogées ayant écouté au moins une fois la radio la veille du jour où ils ont été contactés.

On s'intéresse dans cette présentation à une entreprise confectionnant de la sauce tomate destinée à venir accompagner des bons plats de pâtes. En particulier on étudie l'évolution du prix des tomates qu'elle achète.

Afin de pouvoir donner un prix prévisionnel à une certaine date on souhaite modéliser l'évolution du prix à partir de deux composantes : l'une donnant la tendance de cette série chronologique sur un certain laps de temps assez long et l'autre étudiant une saisonnalité éventuellement observée sur la série. Ici les données étant mensuelles on va chercher à quantifier par un coefficient saisonnier S associé à chaque mois la façon dont le prix se positionne par rapport à la tendance T constatée. En s'appuyant sur un modèle multiplicatif, on modélise donc ceci par la relation suivante :

$$Prix = T \times S$$

La première étape de la méthode utilisée consiste à lisser la série, c'est-à-dire à faire passer une courbe entre les creux et les pics saisonniers. Ceci est établi par un calcul de moyenne mobile et cela fournit une première version de la tendance T (mais qui ne pourra pas être exploitée pour la prévision). La moyenne mobile est attribuée à une date qui doit être parfaitement centrale par rapport aux dates à partir desquelles elle est calculée et elle est calculée sur l'ensemble d'une année de sorte à isoler l'influence d'une saisonnalité éventuelle. Pour cela on envisage deux cas :

— Si le nombre de dates exploitées est impair, un seul calcul de moyenne est nécessaire. Par exemple si une série est systématiquement mesurée sur les mois de janvier à septembre uniquement on peut effectuer les calculs suivants :

$$MM9(t = Mai_{a1}) = \frac{Janv_{a1} + \cdots + Sept_{a1}}{9} = Moy(Janv_{a1} : Sept_{a1})$$
$$MM9(t = Juillet_{a1}) = Moy(Mar_{a1} : Fev_{a2})$$

— Si le nombre de dates exploitées est pair, on doit avoir recours à une moyenne de deux moyennes, établies toujours sur un an mais avec un décalage d'une unité de temps entre les deux, pour avoir une date d'attribution qui soit parfaitement centrale par rapport à l'ensemble des dates utilisées. Ainsi si les relevés sont trimestriels et se font sur l'ensemble des quatre trimestres d'une année, on peut par exemple effectuer les calculs suivants :

$$MM4(t = T3_{a1}) = \frac{moyenne(T1_{a1} : T4_{a1}) + moyenne(T2_{a1} : T1_{a2})}{2}$$
$$MM4(t = T1_{a2}) = \frac{moyenne(T3_{a1} : T2_{a2}) + moyenne(T4_{a1} : T3_{a2})}{2}$$

La deuxième étape consiste en une première estimation des coefficients saisonniers en étudiant pour chaque date la position de la série brute (le prix ici) relativement à la moyenne mobile. Pour cela on effectue simplement le calcul suivant :

$$S_1(date) = \frac{Prix(date)}{MM(date)}$$

.

L'influence saisonnière sur le prix est ici censée être la même chaque année. Les fluctuations observées d'une année sur l'autre sont donc considérées comme des variations exceptionnelles en dehors de la tendance et de ce mouvement saisonnier. Pour lisser ces variations accidentelles et disposer d'un seul coefficient saisonnier pour chaque moment de l'année (chaque mois ici), on effectue la moyenne des coefficients saisonniers obtenus sur les différentes années sur ce même mois. On obtient donc ici une deuxième série de seulement douze coefficients saisonniers S_2 (colonne `coef_sais` dans le tableau représenté ci-après).

Le mouvement saisonnier ne devant pas avoir d'influence sur l'évolution de la série sur le long terme, on va exiger de plus à ce que sa moyenne sur un an (i.e. la moyenne des 12 coefficients mensuels ici) soit parfaitement égale à 1. Pour cela le calcul suivant devra être systématiquement exécuté à partir de la série des coefficients $S2$ pour les corriger et obtenir la série des coefficients $S3$ (ou `coef_sais_corr` dans le tableau ci-après) dont on pourra vérifier que la moyenne est bien égale à 1 :

$$S3(\text{Moment de l'année}) = \frac{S2(\text{Moment de l'année})}{\text{moyenne des } S2 \text{ sur un an}}$$

mois	coef_sais	coef_sais_corr
1	1.039864638	1.0398217477
2	1.0323970...	1.0323544447
3	1.0978984...	1.0978531763
4	1.1972103...	1.1971610177
5	1.0960780...	1.0960328615
6	0.9342826...	0.9342440694
7	0.8890482...	0.8890115975
8	0.8357613...	0.8357269026
9	0.8990677...	0.8990306856
10	1.0047385...	1.0046971435
11	0.9802879...	0.9802474682
12	0.9938598...	0.9938188854

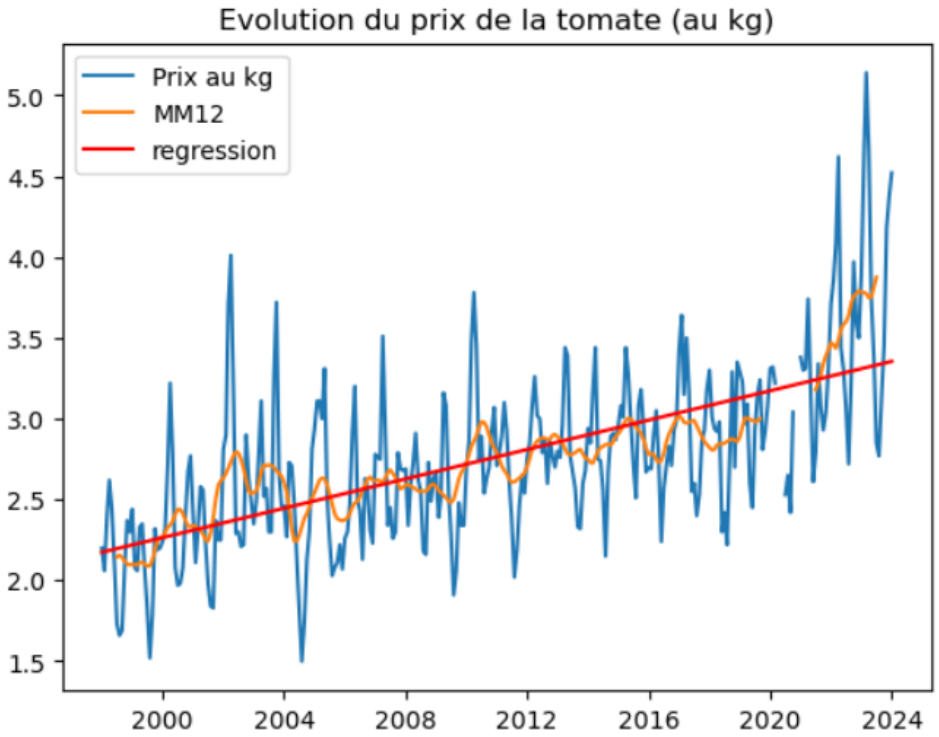
Une fois cette série de coefficients saisonniers obtenue, notre objectif est de pouvoir l'exploiter pour effectuer une prévision du prix à une date ultérieure. Pour cela il faut modéliser la tendance T comme une fonction du temps, et cela va se faire par défaut par l'utilisation d'un modèle de régression linéaire entre la série des prix relevés et une série numérique représentant les dates (soit

par exemple en utilisant le format numpy `datetime64` ce qui a été effectué dans cette présentation soit en fractionnant manuellement le temps en fonction de la périodicité des relevés ce qui sera effectué dans les fonctions à implémenter). En procédant ainsi on obtient la droite représentée ci-dessous dont l'expression est la suivante :

$$T_{prix} = 1.439 \times 10^{-18} \times Date + 0.9024$$

On peut par exemple s'en servir pour effectuer une prévision pour août 2024 en appliquant le modèle construit (nb : au regard d'une tendance à plus court terme on pourrait aussi chercher à établir une tendance à partir d'un mouvement de hausse des prix sensible depuis 2021) :

$$\begin{aligned} Prvision\ Prix(aout\ 2024) =& T_{prix}(aout\ 2024) \times S(aout) \\ =& (1.439 \times 10^{-18} \times datetime(01-08-2024) + 0.9024) \times 0.835727 \end{aligned}$$



Appliquez les méthodes décrites précédemment pour répondre aux questions suivantes à partir des données proposées dans le fichier `audience_cumulee_radio.csv`. Ces données sont trimestrielles et comme dans l'exemple précédent on peut observer l'absence de donnée sur une période de confinement sur le deuxième trimestre de 2020. Pour éviter que ça ne biaise certains calculs on évitera de calculer une moyenne mobile lorsque celle-ci devrait exploiter une donnée manquante. Pour cela vous pourrez utiliser la syntaxe suivante : `mean(skipna=False)`.

Q. 2.1. Implémenter une fonction `decomp_serie()` renvoyant la série représentant la moyenne mobile avec l'indexation par défaut de `pandas` et la série représentant les coefficients saisonniers indexée par les libellés des trimestres.

Q. 2.2. Implémenter une fonction `prevision(annee_prev,trim_prev,serie_coef_sais)` renvoyant la prévision effectuée pour l'année et le numéro de trimestre passés en arguments. La prévision est établie à partir de la tendance modélisée par régression linéaire sur la période étudiée dans le `dataframe data_audience` importé dans le fichier python et le volume d'audience apparaissant dans sa série brute dans la colonne `AC`. Pour cela il faudra toutefois exclure les dates sans relevé disponibles en exploitant par exemples les méthodes `np.logical_not` et `np.isnan`.

Cela n'est pas demandé ici mais pour indication voici le graphique donnant l'évolution de l'audience sur la période étudiée :