TD-TP3
Corrélation entre 2 variables statistiques

• Partie TD (sur papier)

Exercice 1 - *L'âge en fonction du pourcentage de cheveux blancs*

On a vu en cours qu'on pouvait trouver une droite qui permet d'obtenir l'âge approximatif d'un enfant en fonction de sa taille.

On va voir maintenant comment trouver l'âge d'un adulte en fonction du pourcentage de cheveux blancs, à partir des données suivantes :

	Adulte A	Adulte B	Adulte C	Adulte D	Adulte E
% cheveux blancs	0.01	0.18	0.42	0.7	0.99
Âge	21	39	45	54	65

1. Pour les 2 variables statistiques observées : quelle est la population ? quel est l'effectif total ?
2. Faire un graphique représentant ces données, avec le pourcentage de cheveux blancs en abscisse et l'âge en ordonnée. Les variables vous semblent-elles corrélées ?
3. Calculer la covariance de ces 2 variables, puis leur coefficient de corrélation. Sont-elles faiblement ou fortement corrélées ?
4. Calculer les coefficients a et b de la droite $y = ax + b$ d'ajustement linéaire du pourcentage de cheveux blancs sur l'âge. Tracer cette droite sur le graphique.
5. D'après vos résultats précédents, si l'on vous présente un individu ayant 30% de cheveux blancs, à combien estimeriez-vous son âge ?

Exercice 2

On considère les variables statistiques X et Y suivantes sur la même population P :

- P est l'ensemble d'années $\{2005, 2006, 2007, 2008, 2009\}$
- X est le nombre de films où Nicolas Cage apparait,
- Y est le nombre d'éditrices dans la Harvard Law Review.

Les valeurs de X et Y sont données dans le tableau suivant :
Attention, ce sont bien les valeurs de X et Y qui sont données, il ne s'agit pas de leur tableau statistique.

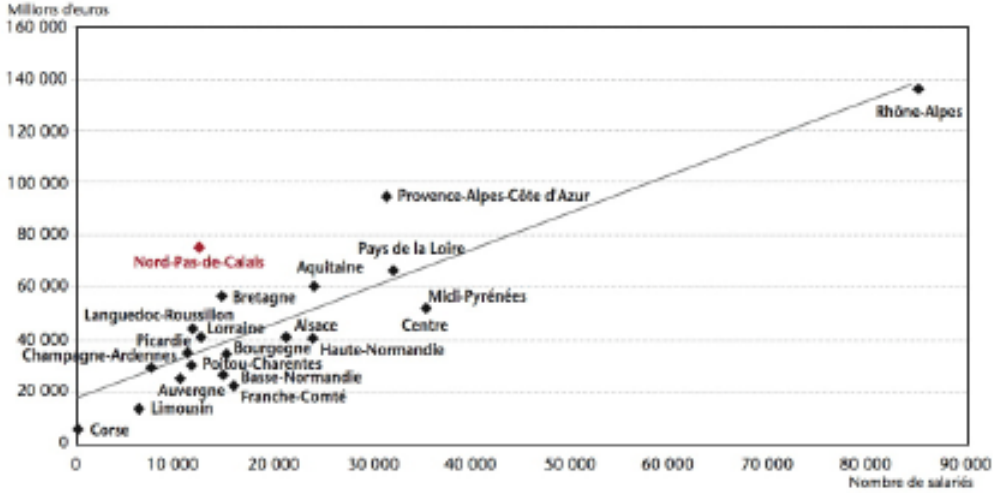
Année	2005	2006	2007	2008	2009
X	2	3	4	2	4
Y	9	14	17	12	18

On se demande (naturellement) si ces 2 séries statistiques sont corrélées.

1. Pour les 2 variables statistiques observées : quelle est la population ? quel est l'effectif total ?
2. Faire un graphique représentant ces données, avec X en abscisse et Y en ordonnée. Les variables vous semblent-elles corrélées ?
3. Calculer le coefficient de corrélation de ces 2 variables. Sont-elles faiblement ou fortement corrélées ?
4. Calculer les coefficients a et b de la droite $y = ax + b$ d'ajustement linéaire de Y en X. Tracer cette droite sur votre graphique.
5. Sachant que Nicolas Cage est apparu dans 6 films en 2023, à combien estimez-vous le nombre d'éditrice dans la Harvard Law Review en 2023 ?

Exercice 3 - *Pour les plus rapides ou pour réviser*

La figure suivante indique, pour les 21 régions françaises, le PIB (y) par région en fonction du nombre d'emplois (x) dans la haute technologie, pour l'année 2000 (source : INSEE Nord-Pas-de-Calais). Le nuage de points, de forme allongée, suggère l'existence d'une relation linéaire (figurée par la droite tracée) entre ces deux variables.



On donne par ailleurs les résultats intermédiaires suivants :

Σx	Σy	Σx^2	Σy^2	Σxy
431200	992600	15078020000	64038160000	29144300000

1. Pour les variables statistiques observées, quelle est la population ? quel est l'effectif total ?
2. Calculer les coefficients a et b, estimations des paramètres de la relation linéaire qu'on cherche à mettre en évidence.
3. La région Nord-Pas-de-Calais ainsi que la région Provence-Alpes-Côte d'Azur sont assez éloignées du modèle obtenu. Selon vous, quelles raisons structurelles propres à ces régions pourraient expliquer cet écart ?

Exercice 4 - *Pour les plus rapides ou pour réviser*

À l'oral d'un examen, chaque candidat est interrogé en première et seconde langue. On note X et Y les notes obtenues dans chaque matière. Les résultats obtenus par 100 candidats sont regroupés dans le tableau ci-dessous :

X \ Y	2	6	10	14	18
2	2	1	0	0	0
6	5	12	3	1	0
10	5	10	25	5	0
12	0	3	12	10	1
16	0	0	1	2	2

1. Écrire les séries statistiques relatives à chaque variable X et Y.
2. Représenter le nuage de points.
3. Calculer les moyennes et écarts-type de X et Y.
4. Calculer la covariance et le coefficient de corrélation. Commenter.

• **Partie TP (sur Python)**

Exercice 5 - *Analyse Macronienne du chômage*

Dans cet exercice, on va étudier la validité de cette hypothèse Macronienne :

Pour cela, on va se demander s'il existe une corrélation entre le nombre de rues par région et le taux de chômage. Vous trouverez les chiffres du chômage en 2016 et du nombre de rues par région dans le fichier TP3.py sur Moodle.

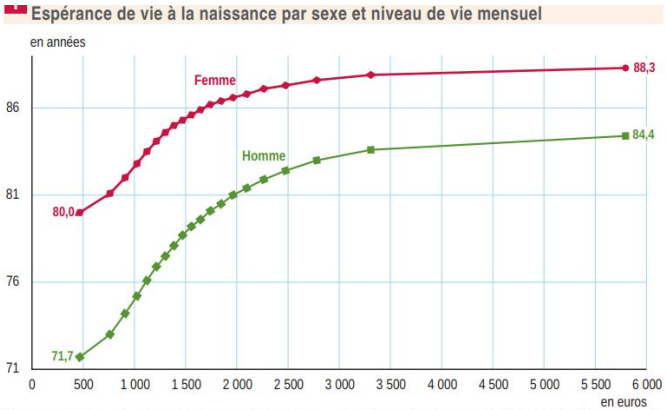
Ces chiffres et cette problématique viennent du projet de statistiques 2019 des anciens étudiants Glenn Anjubault, William Donval, Colas Lefeuvre et Axel Souza.

1. Avec Python, tracer le nuage de points représentant le taux de chômage (Y) en fonction du nombre de rues (X).
   Indication : On pourra se servir du tutoriel d'introduction à Pyplot du site `fr.matplotlib.net`*, notamment de sa partie "Formater le style de votre tracé".*
2. Faire une fonction `Covar(Serie1,Serie2)` qui calcule la covariance de 2 séries statistiques contenues dans des tableaux `Serie1` et `Serie2`.
   Indication : on pourra utiliser la fonction `Moyenne` *du TP1.*
3. Faire une fonction `Correl(Serie1,Serie2)` qui calcule le coefficient de corrélation des 2 séries statistiques contenues dans les tableaux `Serie1` et `Serie2`.
   Indication : on pourra utiliser la fonction `Variance` *du TP1.*
4. Calculer le coefficient de corrélation entre le taux de chômage et le nombre de rues par région. La corrélation est-elle forte ?

5. Créer une fonction `CoeffDroite(Serie1,Serie2)`, qui calcule les coefficients de la droite d'ajustement linéaire de la `Serie2` sur la `Serie1`.
6. Tracer sur votre graphique la droite d'ajustement linéaire du taux de chômage sur le nombre de rues.
7. *Facultatif* - Utiliser la fonction `plt.annotate` pour indiquer le nom de la région à côté des points sur le nuage de points.

Exercice 6

On s'intéresse au graphique suivant, trouvé sur le site de l'*Insee* à l'adresse www.insee.fr/fr/statistiques/3319895 :



Note : en abscisse, chaque point correspond à la moyenne des niveaux de vie mensuels d'un vingtile. Chaque vingtile comprend 5 % de la population.
Lecture : en 2012-2016, parmi les 5 % les plus aisés, dont le niveau de vie moyen est de 5 800 euros par mois, l'espérance de vie à la naissance des hommes est de 84,4 ans.
Champ : France hors Mayotte.
Source : Insee-DGFIP-Cnaf-Cnav-CCMSA, Échantillon démographique permanent.

Ce graphique semble suggérer une corrélation entre le niveau de vie (en revenu mensuel) et l'espérance de vie.
Les points tracés sur le graphique correspondent aux moyennes obtenues pour les *vingtiles* de population, c'est-à-dire :

- le 1e point correspond aux 1/20 les plus pauvres de la population, c'est-à-dire aux 5% de revenus les plus faibles,
- le 5e point correspond aux personnes situées entre les 20% et les 25% les plus pauvres,
- etc...

1. Sur le site de l'Insee, télécharger le tableau excel contenant l'ensemble des données de l'étude. Rentrer ces 3 jeux de données dans 3 tableaux *array* en Python.

Corrélation entre espérance de vie et revenus

2. Aller sur Matplotlib.org dans la rubrique *Tutorials*, et ouvrir le tutoriel *Pyplot tutorial*. Faire/comprendre les 3 premiers graphiques.
3. Tracer le même graphique que l'Insee (on ne demande pas que la mise en forme et les couleurs soient les mêmes !).
4. A l'aide des fonctions faites dans l'exercice précédent, calculer le coefficient de corrélation entre le niveau de vie et l'espérance de vie pour les hommes, puis celui entre le niveau de vie et l'espérance de vie pour les femmes.
   Les corrélations sont-elles fortes ?

Corrélation entre espérance de vie et numéros de vingtiles

Soit `V=np.array([1,2,3,...,19,20])` le tableau contenant les numéros des vingtiles.

5. Tracer un nouveau graphique avec la courbe de l'espérance de vie des hommes en fonction de leur numéro de vingtile, et celle de l'espérance de vie des femmes en fonction de leur numéro de vingtile. Que remarquez-vous ?
6. Calculer le coefficient de corrélation entre le numéro de vingtile et l'espérance de vie pour les hommes, puis celui entre le numéro de vingtile et l'espérance de vie pour les femmes. Qu'observez-vous ? Avez-vous une idée d'explication ?

Facultatif. **Corrélation entre espérance de vie et** log **des revenus**

On s'intéresse maintenant à la corrélation entre l'espérance de vie et le log du niveau de vie.

7. Tracer un nouveau graphique avec en abscisse le log du niveau de vie, et en ordonnée l'espérance de vie des hommes, puis l'espérance de vie des femmes.
8. Calculer le coefficient de corrélation entre le log du niveau de vie et l'espérance de vie pour les hommes, puis celui entre le log du niveau de vie et l'espérance de vie pour les femmes. Qu'observez-vous ?

Exercice 7 - *Sciences étonnantes : Croissance et endettement d'un pays*

Pour cet exercice, notre point de départ est la vidéo suivante, que vous pourrez trouver sur la chaine Youtube "Science Etonnante" de David Louapre :

Regarder cette vidéo, puis accéder aux fichiers de David Louapre, par le lien sous la vidéo. Téléchargez le fichier excel RR.xlsx puis ouvrez dans votre navigateur le fichier RR.ipynb.

1. Dans spyder, importer comme David Louapre la librairie `Pandas`, puis copier et exécuter le code contenu dans les entrées [2], [4] et [5] du fichier RR.ipynb.
2. Utiliser vos fonctions des TP précédents pour calculer le coefficient de corrélation entre la croissance et l'endettement.
   Si besoin, aller dans l'aide de `Pandas` *pour comprendre comment utiliser le tableau* `df`*.*
   La corrélation est-elle faible ? forte ?
3. Toujours à l'aide de vos fonctions des TP précédents, tracer la droite d'ajustement linéaire sur le nuage de points tracé ci-dessus. Obtenez-vous la même droite que David Louapre ?

Ci-contre un extrait du blog de David Louapre, intéressant pour nuancer sa vidéo.

Que celui qui n'a jamais fait d'erreur Excel leur jette la première pierre…

J'espère que le ton de la vidéo fait correctement ressortir le message principal, qui n'est pas de se moquer d'une boulette Excel par des pontes de Harvard, mais surtout de mettre en valeur les structures existantes pour éviter que ces erreurs (humaines, et normales) aient des conséquences :

La revue par les pairs : dont je ne comprends pas toujours pourquoi elle ne s'est pas appliquée dans ce cas là. Apparemment c'était un numéro de la revue regroupant des actes de conférence. Mais pourquoi pas de peer-review ? Et comment un article peut-il alors se prévaloir du prestige de la revue ?

Que celui qui n'a jamais fait d'erreur Excel leur jette la première pierre…

J'espère que le ton de la vidéo fait correctement ressortir le message principal, qui n'est pas de se moquer d'une boulette Excel par des pontes de Harvard, mais surtout de mettre en valeur les structures existantes pour éviter que ces erreurs (humaines, et normales) aient des conséquences :

La revue par les pairs : dont je ne comprends pas toujours pourquoi elle ne s'est pas appliquée dans ce cas là. Apparemment c'était un numéro de la revue regroupant des actes de conférence. Mais pourquoi pas de peer-review ? Et comment un article peut-il alors se prévaloir du prestige de la revue ?

Le partage des données et des codes : il faut reconnaitre que R&R ont quand même eu l'honnêteté intellectuelle de partager leurs données. On n'aurait pas forcément parié dessus, d'autant que la demande émanait de ce qui semble être un « petit » département d'économie voisin. A la fin de cette histoire, c'est quand même David qui a ridiculisé Goliath !

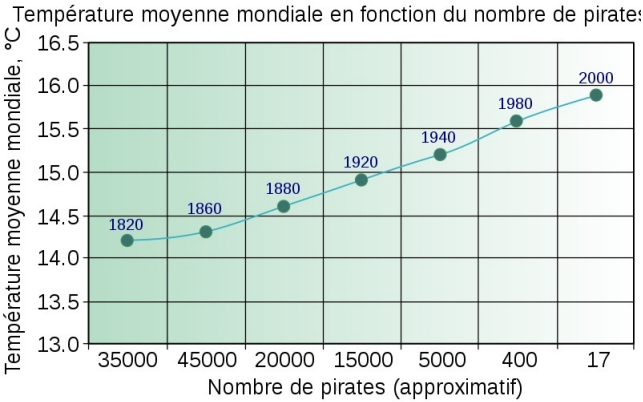
Par ailleurs, sur ce fil Twitter : quelques compléments sérieux d'un macro-économiste sur ce que l'on comprend des liens entre dette et croissance.

Edit du 20/04 : *Suite à plusieurs conversations avec des économistes professionnels, j'ai un peu éclairci cette situation. Avant 2018, le numéro du mois de Mai de l'American Economic Review (AER) était consacré aux proceedings de la conférence de l'American Economic Association. Pour les économistes, il était semble-t-il bien connu que ce numéro de Mai était spécial et différent, sans peer-review formel, avec des papiers plus courts, etc. Donc les « pros » savaient faire la différence en voyant que « c'était dans le numéro du mois de Mai ». Toutefois dès qu'on est un peu extérieur au domaine, si on n'a pas cette connaissance, on ne peut pas deviner qu'une publication « dans le numéro de Mai » n'a pas la même valeur qu'une publication dans les autres numéros de l'AER. Depuis 2018, les choses ont été clarifiées car l'éditeur a décidé de transformer ce numéro spécial en une édition à part entière, avec un vrai nom différent « AEA Papers&Proceedings ». Mais en 2010, quand le R&R a été publié, cette distinction n'avait pas encore été faite.*

Exercice 8 - *Pour les plus rapides*

Selon le pastafarisme (voir dépliant de propagande de cette religion sur Moodle **TP3-Exercice8.pdf**), les pirates sont des « êtres absolument divins », et sont les premiers pastafariens. L'image des pirates présentés comme des « voleurs et des hors-la-loi » est, dans la vision pastafarienne, de la désinformation répandue par les théologiens chrétiens. En réalité, les pirates auraient été des « explorateurs pacifiques répandant la bonne parole » qui distribuaient des friandises aux petits enfants et les pirates modernes n'auraient rien à voir avec les « anciens boucaniers aimant s'amuser ».

Les pastafariens défendent l'idée selon laquelle le réchauffement climatique serait dû à la disparition progressive des pirates : « Nous savons, car nous avons la foi, que le réchauffement climatique est dû au courroux du Monstre en Spaghetti Volant devant la baisse dramatique du nombre de Pirates dans le monde ». Dans le dépliant de propagande **TP3-Exercice8.pdf**, vous trouverez le graphique ci-dessous qui illustre le lien entre la disparition des pirates et le réchauffement climatique, sur les 200 dernières années... sans citer ses sources.



- Chercher en ligne des données sur le nombre de pirates et le réchauffement climatique, et utilisez les notions vues en cours pour confirmer ou infirmer cette théorie.