Projet
Statistiques descriptives à 2 variables : droite de régression

Rappels de statistique :

Soient deux variables statistiques X et Y. Soient x_i les valeurs prises par la variable X (pour i compris entre 1 et n) et y_i les valeurs prises par la variable Y (pour i compris entre 1 et n) .

- Le nuage de points de la série statistique double (X, Y) est l'ensemble des points $M_i(x_i,y_i)$ (où i est compris entre 1 et n).

- Le ***point moyen*** de ce nuage de points est le point G de coordonnées $(\bar{x}, \bar{y})$, où $\bar{x}$ est la moyenne de la variable X, et $\bar{y}$ est la moyenne de la variable Y, c'est à dire pour une série non classée :

$$\bar{x}=\sum_{i=1}^{n}\frac{x_i}{n} \quad \text{et} \quad \bar{y}=\sum_{i=1}^{n}\frac{y_i}{n}$$

- La ***variance*** de la variable statistique X est le réel $var(X)=\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{n}$,

  la variance de Y est le réel $var(Y)=\sum_{i=1}^{n}\frac{(y_i-\bar{y})^2}{n}$

- La ***covariance*** des variables X et Y est le réel $cov(X,Y)=\sum_{i=1}^{n}\frac{(x_i-\bar{x})(y_i-\bar{y})}{n}$

- Le ***coefficient de corrélation linéaire*** est le réel : $r=COR(X,Y)=\frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}$

- La ***droite d'ajustement linéaire par la méthode des moindres carrés*** de ce nuage de points est la droite d'équation : $y=ax+b$ avec $a=\frac{cov(X,Y)}{var(X)}$ et $b=\bar{y}-a.\bar{x}$

  NB : Il n'est pertinent de la tracer que si le coefficient de corrélation est proche de 1 ou de -1.

Dans ce projet, nous souhaitons :

- créer des nuages aléatoires de points et les représenter graphiquement,

- calculer leur coefficient de corrélation,

- décider s'il est pertinent de réaliser un ajustement linéaire (c'est à dire de tracer la droite des moindres carrés) et si oui, tracer la droite des moindres carrés.

1) PARTIE « OUTILS » :

Dans toute cette partie, on considérera des fichiers de texte, contenant, sur chaque ligne, 2 nombres s éparés par un espace, qui pourront être considérés comme l'abscisse x et l'ordonnée y d'un point dans le plan. Chaque ligne d'un tel fichier pourra donc représenter les coordonnées d'un point dans le plan, et un tel fichier pourra représenter un nuage de points, c'est à dire un ensemble de points.

Dans un premier temps, définir les fonctions suivantes :

a) Une fonction nommée cree_fichier_alea(nb, nomfichier) prenant un argument entier et un argument chaîne de caractères. Cette fonction créera sur le disque dur un fichier de texte, qui sera nommé selon le paramètre nomfichier, et qui contiendra nb lignes. Chaque ligne devra être formée de 2 nombres flottants aléatoires entre 0 et 500, séparés par un espace. Un tel fichier pourra ainsi représenter un nuage de points. Pour générer ces abscisses et ordonnées aléatoires, on utilisera une loi uniforme entre 0 et 500 (on importera pour cela le module `random`). Cette fonction ne renverra rien.

b) Une fonction nommée lit_fichier(nomfic) prenant un argument de type chaîne de caractères représentant le nom d'un fichier sur le disque dur. Ce fichier sera supposé contenir sur chaque ligne, 2 nombres séparés par un espace qui représenteront les cooordonnées d'un point d'un nuage de points. La fonction ouvrira et lira ce fichier, puis elle renverra deux listes listeX et listeY. La première liste contiendra toutes les abscisses des points du nuage, c'est à dire tous les nombres de la première colonne du fichier. La deuxième liste contiendra toutes les ordonnées des points du nuage (dans le même ordre que les abscisses) c'est à dire tous les nombres de la 2ème colonne du fichier. Par exemple, si le fichier contient les 3 lignes suivantes :

x1 y1
x2 y2
x3 y3

alors les 2 listes renvoyées par la fonbction seront : [x1,x2,x3] et [y1, y2, y3].

c) Une fonction nommée trace_Nuage(nomf) prenant un argument de type chaîne de caractères, représentant le nom d'un fichier. Ce fichier sera supposé contenir les coordonnées des points d'un nuage sous la forme de lignes de 2 nombres séparés par un espace. Cette fonction appellera tout d'abord la fonction lit_fichier() pour ouvrir le fichier et obtenir les coordonnées des points du nuage, puis elle représentera graphiquement le nuage de points correspondant. Enfin, elle renverra le nombre de points dessinés (on comptabilisera comme des points différents tous les points dont les coordonnées sont fournies par le fichier, même si certains de ces points ont des coordonnées identiques).

d) Une fonction nommée trace_droite(a, b) prenant 2 arguments flottants représentant le coefficient directeur a et l'ordonnée à l'origine b d'une droite, qui déterminera les coordonnéees de 2 points de cette droite afin qu'une ligne reliant ces 2 points soit représentée graphiquement. Cette fonction ne retournera rien.

e) A la fin de cette partie, un jeu d'essais devra tester le bon fonctionnement de ces 4 fonctions.

2) PARTIE « CALCULS STATISTIQUES » :

Dans un 2e temps, écrire les fonctions suivantes :

a) Une fonction nommée moyenne(serie) qui prendra en argument une liste de réels représentant une série statistique.

Cette fonction retournera la moyenne de ces réels.

b) Une fonction nommée variance(serie) qui prendra en argument une liste de réels représentant une série statistique.

Cette fonction retournera la variance de cette série statistique.

Elle devra appeler la fonction moyenne().

c) Une fonction nommée covariance(serieX, serieY) qui prendra en argument 2 listes de réels représentant 2 variables statistiques X et Y.

Cette fonction retournera la covariance de ces deux variables.

Elle devra appeler la fonction moyenne().

d) Une fonction nommée correlation(serieX, serieY) qui prendra en argument 2 listes de réels réprésentant 2 variables statistiques X et Y.

Cette fonction retournera leur coefficient de corrélation linéaire.

Elle devra appeler la fonction variance() et la fonction covariance().

e) Une fonction nommée forteCorrelation(serieX, serieY) qui prendra en argument 2 listes de réels représentant 2 variables statistiques X et Y.

Elle devra décider si X et Y sont fortement liées (corrélées), de la façon suivante :

- elle calculera le coefficient de correlation de X et Y ;

- elle évaluera si ce coefficient est proche de 1 ou de -1 (on considérera qu'un coefficient supérieur à 0,8 ou inférieur à -0,8 est proche de 1 ou de -1).

Enfin elle retournera un booleen égal à :

- True s'il y a une forte corrélation entre X et Y,

- et False si la corrélation est faible.

Elle devra appeler la fonction correlation().

f) Une fonction nommée droite_reg(serieX, serieY) qui calculera les coefficients de la droite de régression de 2 variables statistiques X et Y.

Elle les retournera sous forme d'un tuple (coeff_dir, ord_orig] contenant le coefficient directeur et l'ordonnée à l'origine de la droite de régression.

g) A la fin de ce module, réaliser des tests (jeu d'essais) pour toutes ces fonctions.

3) PROGRAMME PRINCIPAL :

Dans un troisième temps,

a) Créer une fenêtre graphqiue comportant un canevas et trois boutons :

- bouton « Tracer la droite » : lorsque l'on clique dessus, une ligne colorée apparaît dans le canevas.

- bouton « Autre couleur » : si l'on actionne ce bouton, une nouvelle couleur est tirée au hasard dans une série limitée. Cette couleur est celle qui s'appliquera aux tracés de droite suivants,

- bouton « Quitter » : qui sert bien évidemment à terminer l'application en refermant la fenêtre.

b) Mettre à jour votre fenêtre graphique (dimensions, boutons complémentaires, commandes...) et appeler les fonctions des parties 1) et 2) précédentes pour :

- créer un nuage de points aléatoire de 50 points et le représenter graphiquement

- calculer le coefficient de corrélation des points de ce nuage (en prenant comme sérieX, la série des abscisses des points du nuage, et comme série Y la série des ordonnées des points du nuage)

- utiliser la fonction forte_correlation() pour décider s'il est pertinent de tracer la droite de régression de ce nuage ou pas. Tracer cette droite à l'aide du bouton « Tracer la droite » si la corrélation est forte.

c) Recommencer avec le fichier « exemple.txt » fourni.

d) Mettre en place un mode « Dessin » permettant de dessiner un nuage de points à l'aide de la souris, avec par exemple un bouton pour activer ce mode et un bouton pour le désactiver.

Un clic par la suite sur le bouton « Tracer la droite » permettra ainsi de tracer la droite de régression si le coefficient de corrélation est proche de 1 ou de -1.

e) En vous inspirant de l'activité détaillée au lien suivant :
https://www.isnbreizh.fr/nsi/activity/fichierCSVPython/pythonPandasCSV/index.html
écrire une fonction permettant d'extraire du fichier « villes_virgule.csv » fourni, les nombres d'habitants « nb_hab_2010 » et « nb_hab_2012 » inférieurs ou égaux à 500 puis :
- créer un nuage de points avec les n = 100 premières valeurs de « nb_hab_2010 » et « nb_hab_2012 »

- le représenter graphiquement et calculer le coefficient de corrélation des points de ce nuage

- utiliser la fonction forte_correlation() pour décider s'il est pertinent de tracer la droite de régression de ce nuage ou pas et tracer la droite de régression si la corrélation est forte.

f) Modifier le code pour que le nombre « n » de la question précédente soit rentré par l'utilisateur.

g) Recommencer avec un fichier de données réelles de votre choix ; vous pouvez vous inspirez de vos autres cours/TD/TP ou choisir un fichier accessible sur internet, genre :
https://gilles-hunault.leria-info.univ-angers.fr/Datasets/datasets.htm
https://www.stat4decision.com/fr/10-sites-de-reference-open-data/
https://www.data.gouv.fr/fr/organizations/institut-national-de-la-statistique-et-des-etudes-economiques-insee/

Pour aller plus loin : Faites preuve d'initiative en proposant d'autres fonctionnalités et améliorations : optimisation du code, possibilité de choisir une configuration dans une liste déroulante, sauvegarde de plusieurs configurations dans des fichiers...