

MEA3 : Informatique TP #5

Introduction : Éléments de Statistique

Dans ce TP, vous allez analyser des données météorologiques stockées sous forme d'un fichier CSV pour en extraire différentes corrélations. Pour cela, vous aurez besoin de connaître certaines formules de statistiques, présentées ci-dessous :

On suppose deux variables aléatoires X et Y pour lesquelles on dispose d'échantillons de n valeurs.

Espérance	$E(X) = \frac{1}{n} \times \sum_{i=1}^n x_i$
Variance	$V(X) = \frac{1}{n} \times \sum_{i=1}^n (x_i - E(X))^2 = E(X^2) - E(X)^2$
Covariance	$\text{cov}(X, Y) = \frac{1}{n} \times \sum_{i=1}^n ((x_i - E(X)) \times (y_i - E(Y))) = E(XY) - E(X) \times E(Y)$
Corrélation de Pearson	$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{V(X) \times V(Y)}} ; \rho(X, Y) \in [-1; 1]$

La corrélation de Pearson permet d'estimer l'intensité du lien entre X et Y :

$$\begin{aligned}\rho(X, Y) = \pm 1 &\Rightarrow Y = \pm a \times X + b \\ |\rho(X, Y)| \in]0; 1[&\Rightarrow Y = \pm a \times X + b + \varepsilon \\ \rho(X, Y) = 0 &\Rightarrow \text{pas de lien linéaire}\end{aligned}$$

S'il existe un lien linéaire entre X et Y expliquant leur évolution conjointe, celui-ci peut être estimé au moyen d'une régression linéaire. Cette technique permet d'estimer les coefficients de la droite de régression $E(Y) = \beta_0 + \beta_1 \times X + \varepsilon$ par minimisation de la somme des erreurs (ε) au carré. La

solution est alors unique :
$$\begin{cases} \hat{\beta}_1 = \frac{\text{cov}(X, Y)}{V(X)} \\ \hat{\beta}_0 = E(Y) - \hat{\beta}_1 \times E(X) \end{cases}$$

Le fichier [CLimatDepCSV.csv](#) regroupe les données météorologiques annuelles (relevées par Météo France) sur la période 1970-2015.

Ouvrez une première fois le fichier dans un éditeur de texte pour l'analyser. Notez l'ordre des colonnes et les types des données à extraire dans chaque colonne.

Vous pouvez remarquer aux lignes 22 à 29 que les premières valeurs sont à 0. La valeur 0 a été utilisée comme valeur par défaut lorsque la donnée manquait. Lorsque pour une année et un département, toutes les valeurs sont à 0, il faut considérer qu'il n'y a pas de donnée pour cette année et donc ne pas les prendre en compte dans les calculs de statistique.

Partie 1 : Lecture du fichier CSV et création d'une base de données

Dans cette première partie, l'objectif est d'ouvrir le fichier de données *CLimatDepCSV.csv* et d'utiliser son contenu pour construire une base de données structurée.

Pour cela, vous devez vous appuyer sur le fichier d'entête *database.h* qui impose le format de la base de données à l'aide de définition (*typedef*) de structures :

- Le type '*database_t*' est une structure qui contient 4 membres :
 - Un tableau de « régions »
 - Un tableau de « départements »
 - Deux compteurs

database	database_t	{...}
regions	region_t [12]	0x7ffffffe8730
departments	department_t [100]	0x7ffffffeabc0
nb_regions	unsigned int	12
nb_departments	unsigned int	94

- Le type '*region_t*' est une structure qui contient 3 membres :
 - Un tableau de caractère pour stocker le nom de la région
 - Un compteur du nombre de départements associés à cette région
 - Un tableau à deux dimensions (série / année) pour les données qui correspondent aux moyennes annuelles calculées sur l'ensemble des départements associés. Ces données ne sont pas disponibles dans le fichier CSV, il vous appartiendra de les calculer plus tard.

database	database_t	{...}
regions	region_t [12]	0x7ffffffe8730
regions[0]	region_t	{...}
regions[1]	region_t	{...}
regions[2]	region_t	{...}
name	char [40]	0x7ffffffe8d48
nb_departments	unsigned int	6
values	float [4][46]	0x7ffffffe8d74
regions[3]	region_t	{...}
regions[4]	region_t	{...}

```
Name : regions[2]
Details:{name = "Provence-Alpes-Côte d'Azur", '\0' <repeats 13 time
Default:{...}
Decimal:{...}
Hex:{...}
Binary:{...}
Octal:{...}
```

- Enfin, le type '*departement_t*' est une structure qui contient 4 membres :
 - Un tableau de caractère pour stocker le nom du département
 - Le numéro du département (ex. 34)
 - L'index dans la base de données de sa région d'appartenance
 - Un tableau à deux dimensions (série / année) pour les données

▼ database	database_t	{...}
▶ regions	region_t [12]	0x7ffffffe8730
▼ departments	department_t [100]	0x7ffffffeabc0
▶ departments[0]	department_t	{...}
▶ departments[1]	department_t	{...}
▼ departments[2]	department_t	{...}
▶ name	char [30]	0x7ffffffeb1d0
(×)= numero	unsigned int	3
(×)= region_index	unsigned int	0
▶ values	float [4][46]	0x7ffffffeb1f8
▶ departments[3]	department_t	{...}

Name : departments[2]

Details:{name = "ALLIER", '\0' <repeats 23 times>, numero = 3, regi

Default:{...}

Decimal:{...}

Hex:{...}

Binary:{...}

Octal:{...}

En plus des définitions de types, le fichier d'entête *database.h* déclare un certain nombre de fonctions que vous avez à implémenter vous-même, dans un fichier *database.c* qui vous ajouterez au projet. La documentation incluse dans le fichier d'entête sous forme de commentaires vous donne toutes les indications nécessaires pour comprendre ce qui est attendu au niveau de chaque fonction.

La fonction *main()* est donc chargée de :

- Déclarer une variable de type '*database_t*'.
- Procéder à son initialisation (tous les champs à 0) à l'aide d'une fonction *init_database()* à implémenter dans *database.c*.
- Ouvrir en lecture le fichier *CLimatDepCSV.csv*
- Extraire à l'aide d'une boucle chaque ligne du fichier, à tour de rôle, afin de la passer en argument à la fonction *parse_csv_line()* qui devra à son tour s'appuyer sur les fonctions de *database.c* pour effectuer le remplissage de la base de données.
- Fermer le fichier *CLimatDepCSV.csv*

Pour cela, vous aurez besoin de plusieurs fonctions disponibles dans les librairies standards. Le site koor.fr propose de très bonnes documentations et même quelques exemples bien écrits pour certaines fonctions) :

<code>fopen()</code> , <code>fclose()</code>	Pour ouvrir et fermer un fichier sur le disque dur
<code>fgets()</code>	Pour lire une ligne dans un fichier
<code>strtok()</code>	Pour extraire des champs dans une chaîne de caractère à l'aide d'un caractère « séparateur » comme par exemple ';' ;
<code>memset()</code>	Pour initialiser tout un bloc en mémoire à l'aide d'une adresse de départ et d'une taille, qui peut notamment être fournie par un <code>sizeof()</code>
<code>strcpy()</code>	Pour copier une chaîne de caractère dans une autre
<code>strcmp()</code>	Pour évaluer l'égalité entre deux chaînes de caractères
<code>atoi()</code> , <code>atof()</code>	Pour convertir un nombre disponible sous forme de chaîne de caractères en une variable numérique

Conseils : Procédez par étapes et étudiez le comportement des fonctions à l'aide du débogueur. N'essayez pas de tout coder d'un coup.

Partie 2 : Calculs statistiques

Vous devez implémenter les fonctions de calcul suivantes :

- **Moyenne**
- **Variance**
- **Covariance**
- **Corrélation de Pearson**
- **Régression linéaire**

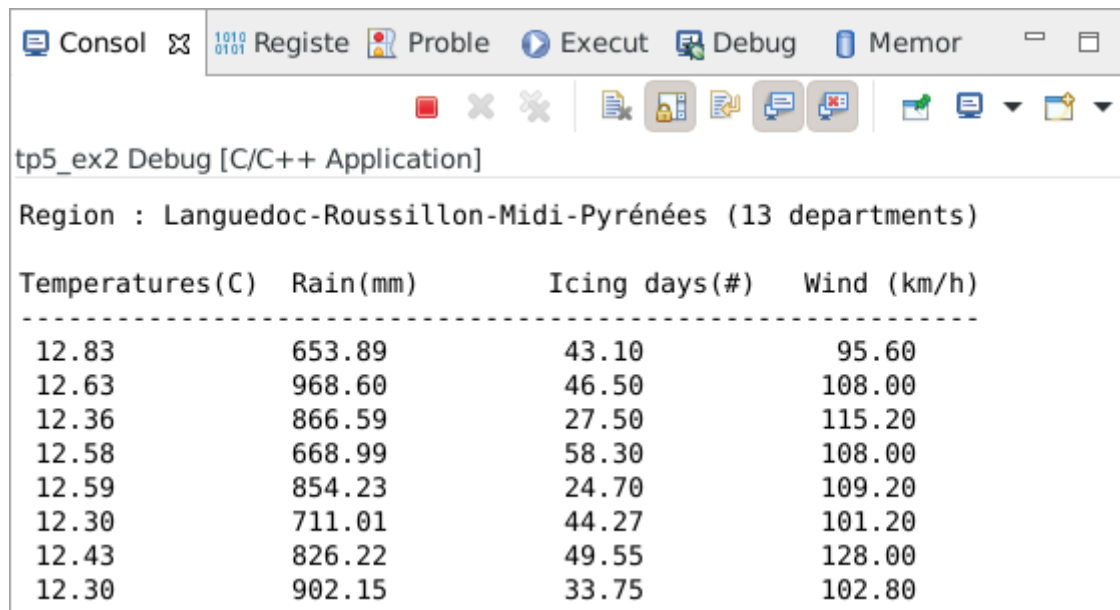
Pour cela, nous allons ici encore développer une petite librairie "`math_stat`" que vous pouvez voir comme une extension de la librairie standard `math` (avec son header `<math.h>`).

Le fichier d'entête `math_stat.h` est imposé. Il définit les entrées/sorties de chacune des fonctions. A vous d'écrire l'implémentation dans un fichier `math_stat.c` que vous ajouterez au projet.

Exercice #1 : Moyenne

Écrivez maintenant une fonction (dans `main.c` par exemple) qui remplit le tableau de données des régions en calculant une moyenne par série et par année à partir des données des départements associés. Attention aux valeurs non renseignées (à 0) à ne pas prendre en compte dans le calcul.

Pour vérifier vos résultats, écrivez également une fonction qui imprime les valeurs d'une région sous la forme montrée ci-dessous. Appuyez-vous sur la feuille de calcul Excel [CLimatDepCSV.xlsx](#) pour vérifier vos résultats.



tp5_ex2 Debug [C/C++ Application]

Region : Languedoc-Roussillon-Midi-Pyrénées (13 departments)

Temperatures(C)	Rain(mm)	Icing days(#)	Wind (km/h)
12.83	653.89	43.10	95.60
12.63	968.60	46.50	108.00
12.36	866.59	27.50	115.20
12.58	668.99	58.30	108.00
12.59	854.23	24.70	109.20
12.30	711.01	44.27	101.20
12.43	826.22	49.55	128.00
12.30	902.15	33.75	102.80

Exercice #2 : Covariance et corrélation de Pearson

Testez vos fonctions de variance, de covariance, et de corrélation en utilisant des séries de données que vous pouvez choisir librement. Le fichier Excel vous donne un exemple en utilisant les températures de la région **Languedoc-Roussillon-Midi-Pyrénées** (variance), avec les températures de la région **Auvergne-Rhône-Alpes** (covariance et corrélation). Sortez vos résultats en console de façon à les vérifier à l'aide de ceux de Excel.

Exercice #3 : Régression linéaire

Testez votre fonction de régression linéaire : Si la tendance se poursuit, quels seront les valeurs (températures, pluviométrie, gelées, vent) en 2050 dans l'Hérault ?

Vérifiez vos résultats à l'aide de la feuille de calcul Excel.