Feuille de TP Python 02

Statistiques descriptives : Traitement de données consignées dans une table Excel

1 Comme d'habitude...

L'objet de ce TP est de mettre en place une méthodologie de traitement de données, données consignées par exemple dans une feuille Excel.

Comme d'habitude, pour démarrer, créez un *répertoire/dossier de travail* `td02` où vous voulez et repérez l'emplacement de ce répertoire. Ouvrez Spyder et dans l'éditeur, tapez (ou faites un copier-coller du .pdf vers l'éditeur) l'entête habituel d'un fichier de script python (*cf.* TP 01) et sauver ce fichier sous le nom `quetelet.py` dans le répertoire de travail `td02`, ce qui nomme le script présent dans l'éditeur.

On y met la `docstring` (entre `"""`) décrivant ce que fait le script et des commentaires (débutant par `#`) signalant les différentes zones du script. .

2 Le fichier de données à traiter

Le fichier [1] `quetelet.csv`, présent dans l'archive du td, contient, pour une série de 66 individus, les valeurs de sexe, taille et poids.

- Placer ce fichier dans votre répertoire de travail.
- Le charger dans LibreOffice Calc ou Micro$oft Excel (ouvrez l'une de ces applications si vous l'avez, sinon passez à la suite) où il sera présenté sous forme de feuille de tableur. Observez la présentation issue de ce logiciel.
- Charger le même fichier dans l'éditeur de Spyder ; observer la structure des informations. Les données sont ici codées sous forme de texte brut : ce que l'on voit est exactement, octet par octet, le contenu du fichier. (Ce n'est pas le cas de la présentation tableur où les données brutes du fichier sont déjà interprétées.)

Le format de ce fichier, dit format CSV (Comma Separated Values), est le format standard d'échange de données statistiques entre logiciels.

C'est un format « Texte », directement lisible et éditable par un éditeur de texte, il a une structure particulière, ce qui permet son interpréation par un logiciel de type tableur (la bonne information ou donnée dans la bonne case).

Dans ce format, la première ligne (souvent) donne les noms des caractères étudiés et *chacune* des lignes suivantes décrit un individu en listant les valeurs des caractères de cet individu et en les séparant par des virgules. Un caractère peut prendre des valeurs numériques (poids, taille) ou des valeurs d'un autre type (sexe).

> *Lors de la récolte de vos données statistiques dans Excel en TIPE, respectez ce format pour constituer vos tableaux (un tableau par feuille Excel))!! Vous pourrez ainsi utiliser les fonctionnalités avancées de Python pour analyser vos données. Excel peut exporter une feuille dans ce format universel.*

Un fichier de format CSV peut-être lu et chargé presque directement dans un objet de type `ndarray`. On peut ensuite effectuer tous les calculs dont on peut avoir besoin pour analyser ces données via Python.

2.1 Charger un fichier .csv dans un tableau

Dans l'onglet `quetelet.py` de l'éditeur, compléter–et sauver en fin d'opération–le script de sorte à obtenir le texte suivant.

```
# -*- coding: utf-8 -*-
"""
quetelet.py : Traitement du fichier quetelet.csv, corrrelations, calcul d'IMC
"""
#Zone (optionnelle) d'importation des modules externes
import numpy as np
import matplotlib.pyplot as plt
#Zone de définition des constantes et fonctions communes
#Script principal
nom_fichier="quetelet.csv"
print("Chargement du fichier de données",nom_fichier)
quetelet = np.genfromtxt(open(nom_fichier,"rb"),delimiter=",",names=True,dtype=None)
print("Fichier chargé")
```

1. La page de l'auteur est `http://irma.math.unistra.fr/~fbertran/enseignement/Statistique_Master1_2011_2.html`

Executez ce script.

1. Afficher dans la console le contenu de l'objet, de type `ndarray`, `quetelet`.
2. Afficher dans la console le contenu des objets `quetelet["sexe"]`, `quetelet["poids"]` et `quetelet["taille"]`
3. Afficher dans la console le contenu des objets `quetelet[0]`, `quetelet[1]`
4. Afficher dans la console le contenu des objets `quetelet[1]["sexe"]`, `quetelet[1]["poids"]`

Qu'en concluez vous sur l'accessibilité des données ? Vérifier qu'il s'agit bien de tableaux Numpy, *i.e.* de type `ndarray`.
Les données portant sur le sexe sont encadrées de guillemets, modifiez le script `quetelet.py` de la façon suivante.

```
quetelet = np.genfromtxt(open(nom_fichier,"rb"),delimiter=",",
                         names=True,converters={0: lambda x : x[1:-1]},dtype=None)
```

Refaites les tests.

2.2 Analyser les données statistiques

En complétant votre script `quetelet.py`,
— donner la moyenne, la variance, l'écart-type [2] du poids et de la taille
— Donner des graphiques montrant les distributions des poids et tailles
— Donner un graphique montrant le nuage $(\text{poids}, \text{taille})$ ainsi que les droites de régression poids/taille et taille/poids.
L'indicateur de masse corporelle (IMC) ou indicateur de QUETELET [3] est défini pour un individu par le rapport

$$\text{imc} = \frac{\text{poids}}{\text{taille}^2}$$

et s'exprime en kg.m^{-2}. L'OMS (Organisation mondiale de la santé interprète cet indice suivant le tableau

IMC (kg.m^{-2}))	Interprétation
moins de 16,5	dénutrition ou famine
16,5 à 18,5	maigreur
18,5 à 25	corpulence normale
25 à 30	surpoids
30 à 35	obésité modérée
35 à 40	obésité sévère
plus de 40	obésité morbide ou massive

TABLE 1 – Interprétation de l'IMC par l'OMS

— Calculer la série statistique des taille^2 (il s'agit des carrés des taille) ;
— Donner un graphique montrant le nuage $(\text{poids}, \text{taille}^2)$ ainsi que la droite de régression poids sur taille^2 ;
— Calculer la série statistique des IMC ;
— en donner moyenne, écart-type, médiane.
— Donner un graphique montrant la distribution de l'IMC sur la population.

2.3 Sélection de données

Pour ce genre de questions, il est probablement maladroit d'étudier une population formée d'hommes et de femmes. La commande suivante fabrique la série statistique ne comportant que les poids des femmes.

```
xpoidsf=np.asarray([quetelet["poids"][i]
          for i in range(len(quetelet["sexe"])) if quetelet["sexe"][i]==b"f"])
```

Il s'agit d'une syntaxe permettant constituer une liste en ne conservant que des objets ayant une certaine propriété. (Définition de liste ou d'ensemble en *compréhension*). Que se passe-t-il si on remplace la chaine `b'f'` par la chaine `'f'` ?
Répondez aux questions du paragraphe précédent en vous limitant aux femmes puis aux hommes. Il peut être intéressant de factoriser le travail fait précédemment dans une fonction `RapportStat(p,t)` où `p` est la série stat. des poids, `t`, celle des tailles.

3 Et ensuite ?

Composer un fichier Excel comportant une feuille consignant, en tableau, quelques caractéristiques physiques de vos camarades (taille, poids, sexe, moyenne de physique en BCPST1,...).
Sauver ce fichier sous format `.csv` et effectuer le même type de travail d'analyse graphique et statistique que celui fait sur fichier `quetelet.csv`. Les résultats de physique de BCPST1 sont-ils corrélés à la taille ?

2. Les éléments de statistiques descriptives à utiliser sont détaillés dans la feuille de TD précédente.
3. Adolphe QUETELET (1796-1874) est un mathématicien, astronome, naturaliste et statisticien belge