TP n° 5 : Statistiques univariées et bivariées

Ce TP est constitué de deux parties complémentaires. Dans un premier temps on introduit les notions au programme pour l'étude de statististiques univariées ou bivariées ainsi que les méthodes de calcul allant avec. La partie "TP" proprement dite intervient dans un deuxième temps.

1 Statistiques univariées

1.1 Vocabulaire et notations

On considère un ensemble donné Ω que l'on appelle *population*, dont on appelle les éléments des *individus*.
On va étudier l'expression d'un *caractère observé* sur les individus, dont on dira que la valeur est la *variable statistique*, notée X.
On dira que X est *discrète* si les valeurs prises par X sont isolées les unes des autres, et *continue* si elles peuvent parcourir un intervalle de $\mathbb{R}$.

Par exemple, si on étudie le taux de chômages dans les pays du monde, on aura :
- ⋆ la population Ω étudiée est l'ensemble des pays du monde,
- ⋆ le caractère observé X est le taux de chômage,
- ⋆ le caractère X est continu, car il parcourt $[0, 1]$, intervalle de $\mathbb{R}$.

Comme il est en général difficile d'observer l'expression de X sur toute la population, on étudie seulement les individus d'une partie de Ω, que l'on notera E et que l'on appelle *échantillon observé*. L'entier $n = Card(E)$ est appelé la *taille* de l'échantillon. A chaque individu de l'échantillon correspond une valeur de X. On obtient alors la *série statistique de l'échantillon*. On notera $X = (x_1, \ldots, x_n)$

1.2 Description d'une série graphique

a) Cas discret

Dans le cas discret, pour étudier X il suffit de connaître l'ensemble $\{a_1, ..., a_m\}$ des valeurs prises par X, appelées *modalités* de X et pour chaque a_j le nombre n_j d'individu correspondant, appelé *effectif* de la modalité. On a donc en particulier $\sum n_j = n$.

Définition 1.1 *On appelle* fréquence *de la valeur a_j le quotient de l'effectif n_j par l'effectif total n. On la notera f_j. En pratique, c'est le taux de la population dont le caractère X prend la valeur a_j.*
Définition 1.2 *On appelle* fréquence cumulée *de a_j le réel $p_j = \sum_{k \leq j} f_k$. C'est le taux de la population dont le caractère prend une valeur inférieure ou égale à a_j.*

b) Cas continu

Dans le cas d'une variable X continue (ou même si le nombre de valeurs prises par X est trop grand), on regroupe les valeurs x_i prises par X appartenant à un même intervalle, que l'on appellera *classe*. Ainsi si l'échantillon observé de X est :

$$[1.2, 0.1, 1.4, 4.5, 3.2, 0.3, 0.6]$$

et si on choisit les classes $[0, 1[$, $[1, 2[$, $[2, 3[$,$[3, 4[$, $[4, 5[$ alors les classes ont respectivement été observées 3 fois, 2 fois, 0 fois, 1 fois et 1 fois.

1.3 Indicateurs de position

a) Moyenne

Soit X échantillon d'une variable statistique discrète. On appelle *moyenne* de X le réel :

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \ ;$$

qui se calcule grâce aux fréquences par :

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{m} n_j a_j = \sum f_j \ a_j \ .$$

Dans le cas d'une variable continue, on remplace les a_j par les centres des classes c_j.

b) Médiane et quantiles

Soit X un échantillon d'une variable statistique discrète. On appelle *médiane* de X la modalité x_i telle qu'il y ait autant de valeurs de X supérieures ou égales à x_i que de valeurs inférieures ou égales. Autrement dit si les valeurs de X ordonnées sont

$$v_1 \leq v_2 \leq ... \leq v_n$$

alors si n est impair égal à $2k+1$, la médiane est v_{k+1}. Si n est pair égal à $2k$ on définit la médiane comme égale à :

$$\frac{v_k + v_{k+1}}{2}.$$

On peut généraliser la notion en ne prenant pas seulement une valeur coupant en deux masses égales les valeurs mais en coupant en quatre masses (quartiles), dix masses (déciles) ou même cent masses (centiles).

1.4 Indicateurs de dispersion

a) Variance empirique

Soit X échantillon d'une variable statistique discrète. On appelle *variance empirique* de X le réel positif :

$$V = \frac{1}{n}\sum n_j(a_j - \overline{x})^2 = \sum f_j(a_j - \overline{x})^2$$

On peut également le calculer par :

$$V = \left(\sum f_j \ a_j^2\right) - \overline{x}^2.$$

On appelle enfin *écart type empirique* de X le réel

$$\sigma = \sqrt{V}.$$

On peut étendre ces définitions en prenant les centres des classes dans le cas continu.

b) Etendue et intervalle interquartile

On appelle *étendue* d'une série statistique la distance entre la plus grande et la plus petite valeur de la série.
On appelle *intervalle interquartile* l'intervalle $]Q_1, Q_3[$ si Q_1, Q_2 et Q_3 sont les quartiles de la série. On appelle *écart interquartile* le réel positif $Q_3 - Q_1$.

1.5 Commandes pythons

Pour obtenir ces données avec python, on utilise numpy :

```
import numpy as np

X = [1,2,1,5,1,5,3,6,9,3]

print(np.mean(X))
print(np.var(X))
print(np.std(X))
print(np.median(X))
```

2 Statistiques bivariées

Dans une population, on peut considérer deux caractères plutôt qu'un seul et étudier le lien entre les deux. Si on note X et Y les deux variables statistiques correspondant à deux caractères étudiés, on cherchera à étudier le couple $Z = (X, Y)$. En particulier on regardera souvent si X influe sur Y, auquel cas on appellera X la *variable explicative* et Y la *variable à expliquer*

2.1 Représentation

On représente les deux séries statistiques sous forme de points dont les abscisses sont les valeurs de X et les ordonnées les valeurs de Y. On obtient ainsi un *nuage de points*. Exemple :

```
import matplotlib.pyplot as plt

X=[1,2,3,5,7]
Y =[4,8,10,16,23]
plt.plot(X,Y,'.')
```

On calcule souvent le *point moyen* du nuage, dont les coordonnées sont $\overline{x}$ et $\overline{y}$ les moyennes empiriques de X et Y. On peut le rajouter au graphique, mais il vaut mieux le distinguer des autres en utilisant un autre style.

On peut déjà mener une première étude qualitative ne considérant le nuage de points : points aberrants, alignement selon une courbe ...

2.2 Modèle de régression

On cherche souvent à déterminer un *modèle de régression*, autrement dit une fonction f telle que l'on approximativement :

$$Y \simeq f(X)$$

On considérera que le modèle déterminé est satisfaisant si l'erreur $\epsilon = Y - f(X)$ est une variable d'espérance nulle et que le nuage de points (X, ϵ) est uniformément réparti autour de 0.

Si on trouve un tel modèle, on dit que X et Y sont *corrélées*. Plus ϵ est petit, plus la corrélation trouvée est forte.

On pourra tester la corrélation en calculant la *covariance empirique* de (X, Y). On la définit par :

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

et qui peut encore se calculer avec l'expression :

$$cov(X,Y) = \frac{1}{n}\left(\sum x_i y_i\right) - \overline{xy}\ .$$

2.3 Régression linéaire

L'exemple le plus classique de modèle est le modèle de régression linéaire, autrement dit on cherche deux constantes a et b telles que :

$$Y \simeq aX + b$$

On a alors :

Propriété 2.1 *La droite approchant de la meilleure façon au sens des moindres carrés le nuage de points est la droite d'équation $y = ax+b$ avec :*

$$a = \frac{cov(X,Y)}{V(X)}$$

et

$$b = \overline{y} - a\overline{x}$$

On l'appelle droite de régression *des séries.*

Exemple 2.1 *1. Écrire une fonction python donnant la covariance empirique de deux séries statistiques.*

2. ; Écrire une fonction donnant le coefficient de corrélation empirique de deux séries statistiques.

3. Tracer la droite de régression associée aux points de l'exemple du 2.1.

La trame des programmes est dans le fichier TP_ Stats.py

Comme on l'a vu dans l'étude des couples de variables aléatoires en probabilités, pour s'assurer que l'utilisation d'un modèle de régression linéaire est bien approprié, on peu calculer le coefficient de corrélation linéaire :

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma(X)\sigma(Y)}$$

plus ce coefficient est proche de 1 ou -1, plus la corrélation linéaire est acceptable.

3 Exemple

On se propose d'étudier un exemple "réel", à partir de données concernant la France métropolitaine et l'évolution de sa consommation d'énergie.

On considère les tableaux suivants :

A	1950	1956	1960	1965	1969	1973	1979	1985	1990	2000
W	63	85	90	115	145	180	193	202	229	269
G	30	41	50	66	80	100	120	132	154	188

A	2005	2006	2008	2009	2010	2011	2012
W	277	276	273	261	263	265	259
G	203	208	213	206	210	214	214

où A est l'année, W la consommation d'énergie primaire en TEP et G le PIB en volume.

1. Tracer le nuage de points associé à (W, G).
2. Déterminer le coefficient de corrélation de W et G.
3. Séparer ensuite les données en W_1 et G_1 d'une part, concernant la période 1950-2000 et d'autre part W_2 et G_2, concernant la période 2005-2012. Qu'observe-t-on ?
4. Déterminer la droite de régression en ne considérant que les années de 1950 à 2000. Tracer la droite. Comment interpréter ce graphique ?