

1 Moyenne d'une série statistique

1.1 Définition mathématique

Définitions.

Dans tout ce qui suit on note :

$$X = (x_1, x_2, x_3, \dots, x_n)$$

une **série statistique**, c'est à dire une suite finie de réels.

On appelle **effectif** de la série statistique, son nombre d'éléments n .

La **moyenne** d'une série statistique X d'effectif n est alors :

$$\bar{X} = \frac{1}{n} \sum_{k=1}^n x_k$$

1.2 Implantation informatique

En Python, une série statistique pourra se donner à l'aide d'une liste de nombres entiers ou flottants. On peut obtenir une liste aléatoire de nombres 100 nombres dans $[0, 10[$ par les instructions suivantes :

```
from random import random
```

```
X = [ 10 * random() for k in range(100) ]
```

☞ La fonction `random()` du module de même nom `random` renvoie un nombre aléatoire entre 0 (inclu) et 1 (exclu).

Ainsi définie, l'effectif de la série statistique s'obtient facilement à l'aide de la fonction `len()` qui renvoie le nombre d'éléments d'une liste (ou séquence) :

```
print("Effectif =", len(X))
```

Exercice 1.

- Écrire une fonction `moyenne(X)` qui prend en paramètre une liste de nombres X et qui renvoie sa moyenne.
- Obtenir cinq listes aléatoires de $n = 10, 100, 1\,000, 10\,000, 100\,000$ nombres dans l'intervalle $[0, 10[$, et calculer pour chacune sa moyenne. Vérifier que lorsque n augmente, la moyenne a tendance à se rapprocher de 5.

2 Variance d'une série statistique

2.1 Définition mathématique

Définition.

Soit $X = (x_i)_{1 \leq i \leq n}$ une série statistique d'effectif n , et $\bar{X}$ sa moyenne. Soit $(X - \bar{X})^2$ la série statistique d'effectif n dont le terme de rang i est :

$$(x_i - \bar{X})^2$$

La **variance** de la série X , notée $V(X)$ ou σ_X^2 , est la moyenne de la série $(X - \bar{X})^2$.

La variance d'une série statistique est la moyenne de la série des carrés des écarts à la moyenne. Elle mesure la **dispersion** des valeurs de la série autour de leur valeur moyenne. Elle a notamment pour propriétés :

$$- V(X) \geq 0,$$

$$- V(X) = 0 \implies \forall i \in [[1, n]], x_i = \bar{X}.$$

Définition.

L'**écart-type** de la série statistique X , noté σ_X , est défini par : $\sigma_X = \sqrt{V(X)}$.

Exercice 2.

- Soit $Y = (X - \bar{X})$; appliquer la fonction `moyenne()` écrite en 1 sur des série statistiques quelconques, et vérifier que l'on a toujours : $\bar{Y} = 0$.
- Le démontrer par le calcul.
- Écrire une fonction `variance()` prenant en paramètre une liste de nombres entiers ou flottants et qui renvoie la variance de la série statistique.
- Vérifier sur des séries aléatoires que $V(X) > 0$.

3 Minimum et maximum d'une série statistique

3.1 Définition mathématique

Définition.

Soit $X = (x_i)_{1 \leq i \leq n}$ une série statistique d'effectif n .

Son minimum, noté $\min(X)$ est $\min \{x_i | i \in [[1, n]]\}$. Son maximum, noté $\max(X)$ est $\max \{x_i | i \in [[1, n]]\}$.

3.2 Implantation informatique

Déterminer le minimum (ou maximum) d'une liste X de nombres est une opération très importante. L'algorithme est le suivant (calcul de minimum) :

Calcul du minimum dans X :

- Initialement m est le premier élément de X .
- Parcourir X du 2ème au dernier élément :
 - Pour chaque x parcouru :
 - Si $x < m$ alors m devient x .
- À la fin du parcours renvoyer m .

Exercice 3.

- a) Appliquer l'algorithme ci-dessus pour écrire une fonction `minimum(X)` prenant en paramètre une liste X de nombres entiers ou flottants, et qui renvoie son minimum.
- b) Adapter l'algorithme ci-dessus pour écrire une fonction `maximum(X)` prenant en paramètre une liste X de nombres entiers ou flottants, et qui renvoie son maximum.
- c) Tester les deux fonctions sur des listes aléatoires.

4 Droite de régression linéaire

La droite de régression linéaire est un outil statistique qui revêt un très grand intérêt en sciences expérimentales. Si lors d'une expérience X et Y sont deux grandeurs numériques entre lesquelles on souhaite établir une corrélation, une loi exprimant Y en fonction de X , ou tout du moins une tendance, on répète l'expérience et on mesure les valeurs prises par X et Y ; mais leurs valeurs sont soumises à des fluctuations en raison de l'imprécision des mesures ou d'aléas. La régression linéaire est l'outil statistique qui permettra dans une telle situation la recherche d'une corrélation indépendante des fluctuations aléatoires.

4.1 Définitions mathématiques

Définitions.

Soit X et Y deux séries statistiques (numériques) ayant même effectif n :

$$X = (x_1, x_2, x_3, \dots, x_n) \quad ; \quad Y = (y_1, y_2, y_3, \dots, y_n)$$

Le **nuage de points** d'abscisses dans X et d'ordonnées dans Y est l'ensemble des points $M(x_i, y_i)$ du plan rapporté à un repère orthonormé, $i \in \llbracket 1, n \rrbracket$. C'est un ensemble de n points du plan.

La droite approchant le mieux le nuage de points, au sens des moindres carrés, est la **droite de régression linéaire**. C'est la droite d'équation

$$Y = a.X + b$$

$$\text{avec} \quad a = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{X})^2} \quad \text{et} \quad b = \bar{Y} - a.\bar{X}$$

où $\bar{X}$ et $\bar{Y}$ désignent les moyennes des séries X et Y .

Le coefficient de corrélation est défini comme :

$$r = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{X})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{Y})^2}}$$

Alors $r \in [-1, 1]$ et la droite de régression linéaire approche d'autant mieux le nuage de point que $|r|$ est proche de 1.

- En notant pour deux séries statistiques $X = (x_1, x_2, x_3, \dots, x_n)$ et $Y = (y_1, y_2, y_3, \dots, y_n)$, de même effectif n :

- La **variance** de X , $V(X) = \overline{(X - \bar{X})^2}$, soit :

$$V(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2$$

- L'**écart-type** de X , $\sigma_X = \sqrt{V(X)}$, soit :

$$\sigma_X = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2}$$

- La **covariance** de X et Y , $cov(X, Y) = \overline{(X - \bar{X})(Y - \bar{Y})}$, soit :

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

la **droite de régression linéaire** $Y = a.X + b$ est alors définie par :

$$a = \frac{cov(X, Y)}{V(X)} \quad ; \quad b = \bar{Y} - a.\bar{X}$$

et son **coefficient de corrélation** est :

$$r = \frac{cov(X, Y)}{\sigma_X \cdot \sigma_Y}$$

4.2 Implantation informatique

Passons à l'implantation En Python.

Exercice 4. On utilisera les fonctions `moyenne()` et `variance()` écrites au dessus.

- Écrire une fonction `covariance(X, Y)` prenant en paramètre deux listes de nombres de même longueur, et renvoyant la covariance de X et Y .
- Écrire une fonction `droiteRegression(X, Y)` prenant en paramètre deux listes de nombres de même longueur, et renvoyant le couple `[a, b]` des coefficients a et b de la droite de régression linéaire.
- Écrire une fonction `coefCorrelation(X, Y)` prenant en paramètre deux listes de nombres de même longueur, et renvoyant leur coefficient de corrélation.

4.3 Application

En laboratoire on a mesuré l'évolution de la concentration d'un réactif lors d'une réaction chimique. Les différents temps de mesure et concentrations mesurées figurent dans le tableau ci-contre.

On souhaite modéliser l'évolution de la concentration de ce réactif, et pour cela nous allons confronter deux modèles cinétiques, celui d'une réaction chimique d'ordre 1 avec celui d'une réaction chimique d'ordre 2.

Dans ces modèles la vitesse d'accroissement $C'(t)$ au temps t de la concentration d'un produit est proportionnelle respectivement à la concentration $C(t)$ au temps t ou à son carré $C(t)^2$, c'est à dire :

Temps (s)	Concentration (mol/l)
0	34,97
3	17,52
5	13,23
7	10,13
10	8,16
12	6,55
15	5,78
18	5,22
20	4,66
23	4,15
25	3,86

Ordre 1 : $\frac{d}{dt}C(t) = -\lambda C(t)$

Ordre 2 : $\frac{d}{dt}C(t) = -\lambda C(t)^2$

On admettra ici (et l'on montrera en cours de Mathématiques) que les solutions à ces équations différentielles sont de la forme (respectivement à l'ordre 1 et à l'ordre 2) :

$$C(t) = C_0 \exp(-\lambda t) \quad ; \quad C(t) = \frac{C_0}{1 - \lambda C_0 t}$$

pour C_0 une constante réelle.

Exercice 5.

- a) Montrer que si la réaction chimique est d'ordre 1, alors en posant $Y(t) = \ln(C(t))$, la fonction $Y(t)$ est affine :

$$Y(t) = -\lambda \cdot t + \ln(C_0).$$

- b) Montrer que si la réaction chimique est d'ordre 2, alors en posant $Z(t) = 1/C(t)$, la fonction $Z(t)$ est affine :

$$Z(t) = -\lambda \cdot t + \frac{1}{C_0}.$$

- Déclarer des listes `T` et `C` contenant les temps de mesure et les concentrations mesurées données dans le tableau ci-dessus.
- En déduire les deux listes `Y` des $\ln$ des concentrations, et `Z` des inverses des concentrations.
- Appliquer une régression linéaire avec les listes `T`, `Y` et `Z`, et à l'aide des fonctions établies à l'exercice 1, et déterminer le modèle à retenir pour cette réaction chimique.
- Tracer dans un graphique l'évolution de la concentration en fonction du temps donnée par le modèle retenu, et le nuage de points des concentrations relevées.

☞ Avec les options `marker='o'`, `linestyle='None'` et `color = 'black'`, l'instruction `plot(T,C,marker='o',linestyle=None, color = 'black')` trace une nuage de points, tous noirs.

On peut aussi utiliser : `plot(T,C,'ok')` (où le "o" remplace `marker='o'`, `linestyle='None'` et le "k" remplace `color = 'black'`).

Voir l'aide en ligne pour l'utilisation de `plot()` :

https://matplotlib.org/api/_as_gen/matplotlib.pyplot.plot.html