Plaquette 2 – Distance entre la moyenne d'un échantillon et l'espérance

Échantillonnage — Première spé maths

Corrections détaillées

Boîte à outils

Distance à l'espérance. Pour un échantillon de moyenne m, l'écart avec l'espérance est $|m - \mu|$; en Python, *abs(m - mu)*.

Ce qu'on observe par simulation (pour N échantillons de taille n) :

— la moyenne des N moyennes est très proche de μ ;

— les moyennes se dispersent autour de μ, avec un écart type proche de

$$\frac{\sigma}{\sqrt{n}}$$

— multiplier n par 4 divise à peu près par 2 la distance typique ; multiplier n par 100 la divise par 10 ;

— à taille égale, une variable deux fois plus dispersée donne des moyennes deux fois plus dispersées.

Écart type d'une liste de valeurs $x_1, \dots, x_n$ de moyenne m :

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - m^2}$$

Rappels. Dé équilibré : $\mu = 3{,}5$, $\sigma = \sqrt{\frac{35}{12}} \approx 1{,}708$. Variable de Bernoulli de paramètre p : $\mu = p$, $\sigma = \sqrt{p(1-p)}$.

Correction de l'exercice 1 – Dix distances à l'espérance

a) On soustrait 3,5 et on prend la valeur absolue : 0,06 ; 0,07 ; 0,34 ; 0,37 ; 0,18 ; 0,15 ; 0,51 ; 0,21 ; 0,17 ; 0,07.

b) La plus grande distance est 0,51 (moyenne 4,01). La somme des distances vaut 2,13, d'où une distance moyenne

$$\frac{2{,}13}{10} = 0{,}213$$

c) Les distances strictement inférieures à 0,2 sont 0,06 ; 0,07 ; 0,18 ; 0,15 ; 0,17 ; 0,07 : 6 échantillons sur 10, soit 60 %. Avec des échantillons de 100 lancers, la moyenne s'écarte donc souvent de plus de 0,2 de l'espérance.

Réponse. a) 0,06 ; 0,07 ; 0,34 ; 0,37 ; 0,18 ; 0,15 ; 0,51 ; 0,21 ; 0,17 ; 0,07 ; b) 0,51 ; moyenne 0,213 ; c) 60 %.

Correction de l'exercice 2 – Programmer le calcul des distances

a) La liste contient 10 nombres : pour chacun des 10 échantillons de 100 lancers, la distance entre sa moyenne et l'espérance 3,5. Au total, $10 \times 100 = 1\,000$ lancers sont simulés.

b) La fonction calcule la moyenne des N distances : c'est la distance « typique » entre la moyenne d'un échantillon de taille n et l'espérance.

c) La fonction *max* de Python renvoie le plus grand élément d'une liste :

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite.
Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).

```
def distance_max(N, n):
    return max(distances(N, n))
```

Si l'on ne veut pas utiliser *max*, on parcourt la liste en gardant la plus grande valeur rencontrée :

```
def distance_max(N, n):
    d = distances(N, n)
    grand = d[0]
    for x in d:
        if x > grand:
            grand = x
    return grand
```

Réponse. a) 10 distances, 1 000 lancers ; b) la distance moyenne ; c) *return max(distances(N, n))*.

Correction de l'exercice 3 – Quand n est multiplié par 4

a) $\dfrac{0{,}2739}{0{,}1345} \approx 2{,}04$ et $\dfrac{0{,}1345}{0{,}0689} \approx 1{,}95$: quand n est multiplié par 4, la distance moyenne est à peu près divisée par 2.

b) On trouve $0{,}2739 \times 5 \approx 1{,}37$, puis $0{,}1345 \times 10 \approx 1{,}35$ et $0{,}0689 \times 20 \approx 1{,}38$. Le produit est presque constant : on conjecture que la distance moyenne est proportionnelle à $\dfrac{1}{\sqrt{n}}$,

$$d \approx \frac{1{,}36}{\sqrt{n}}$$

c) Pour $n = 10\,000$, $\sqrt{n} = 100$, d'où une distance moyenne d'environ $\dfrac{1{,}36}{100} \approx 0{,}014$. Pour diviser la distance par 10, il faut un échantillon 100 fois plus grand.

Réponse. a) Par environ 2 ; b) $d\sqrt{n} \approx 1{,}36$, constant : d proportionnelle à $\dfrac{1}{\sqrt{n}}$; c) environ 0,014.

Correction de l'exercice 4 – Écart type des moyennes et σ/√n

a) Pour chaque taille, la moyenne des 500 moyennes est très proche de l'espérance 3,5 : les moyennes d'échantillons se répartissent autour de μ, sans la surestimer ni la sous-estimer systématiquement.

b) Avec $\sigma \approx 1{,}7078$, on obtient

$$\frac{\sigma}{5} \approx 0{,}3416 \qquad \frac{\sigma}{10} \approx 0{,}1708 \qquad \frac{\sigma}{20} \approx 0{,}0854$$

Ces valeurs sont très proches des écarts types observés (0,3436 ; 0,1693 ; 0,0862). On constate que les moyennes des échantillons de taille n ont un écart type d'environ $\dfrac{\sigma}{\sqrt{n}}$: elles sont $\sqrt{n}$ fois moins dispersées que les valeurs de la variable elle-même.

Réponse. a) Toujours proche de 3,5 ; b) 0,3416 ; 0,1708 ; 0,0854, très proches des écarts types observés.

Correction de l'exercice 5 – Deux jeux de même espérance

a) Jeu A : $E = 2 \times 0{,}5 = 1$, $E(X^2) = 4 \times 0{,}5 = 2$, donc $V = 2 - 1 = 1$ et $\sigma_A = 1$. Jeu B : $E = 10 \times 0{,}1 = 1$, $E(X^2) = 100 \times 0{,}1 = 10$, donc $V = 10 - 1 = 9$ et $\sigma_B = 3$. Les deux jeux ont la même espérance, 1 €.

b) $\sqrt{50} \approx 7{,}07$, d'où

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite.
Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).

$$\frac{\sigma_A}{\sqrt{50}} \approx 0{,}141 \qquad \frac{\sigma_B}{\sqrt{50}} \approx 0{,}424$$

Ces valeurs sont proches des écarts types observés (0,143 et 0,455) : à taille égale, les gains moyens du jeu B sont environ trois fois plus dispersés, comme son écart type.

c) Au jeu B : sa variable est plus dispersée ($\sigma_B = 3\sigma_A$), et sur 50 parties son gain moyen s'écarte bien plus souvent de l'espérance 1 € ; dans la simulation, il a varié de 0 à 2,6 €.

Réponse. a) $E = 1$ pour les deux, $\sigma_A = 1$, $\sigma_B = 3$; b) $\approx 0{,}141$ et $\approx 0{,}424$, proches des valeurs observées ; c) le jeu B.

Correction de l'exercice 6 – Taille nécessaire pour une précision donnée

a) On cherche n tel que $\frac{1{,}708}{\sqrt{n}} \leq 0{,}05$. Comme $\sqrt{n} > 0$, cela équivaut à

$$\sqrt{n} \geq \frac{1{,}708}{0{,}05} = 34{,}16 \Leftrightarrow n \geq 34{,}16^2 \approx 1\,166{,}9$$

la fonction carré étant croissante sur $[0\,;+\infty[$. Il faut donc au moins 1 167 lancers.

b) Ici $\sigma = \sqrt{0{,}5 \times 0{,}5} = 0{,}5$. On veut $\frac{0{,}5}{\sqrt{n}} \leq 0{,}01$, soit $\sqrt{n} \geq 50$, c'est-à-dire $n \geq 2\,500$. Pour gagner un facteur 10 en précision, il faut toujours un échantillon 100 fois plus grand.

Réponse. a) $n \geq 1\,167$; b) $n \geq 2\,500$.

Correction de l'exercice 7 – Fréquences d'échantillons

a) $\mu = p = 0{,}3$ et $\sigma = \sqrt{0{,}3 \times 0{,}7} = \sqrt{0{,}21} \approx 0{,}458$.

b) On obtient

$$\frac{0{,}458}{10} \approx 0{,}046$$

C'est, d'après les simulations, l'écart type des fréquences de sondages de 100 personnes : la distance typique entre une fréquence observée et la proportion 0,3 est de l'ordre de 4 à 5 points.

c) Les deux extrêmes sont à 0,15 de 0,3, soit environ $\frac{0{,}15}{0{,}046} \approx 3{,}3$ écarts types. Ce sont les valeurs les plus éloignées parmi 1 000 sondages : de tels écarts sont très rares. La grande majorité des fréquences restent à moins de deux écarts types, soit 0,09, de la proportion réelle.

Réponse. a) $\mu = 0{,}3$, $\sigma \approx 0{,}458$; b) $\approx 0{,}046$, écart type des fréquences ; c) rares, à environ 3,3 écarts types.

Correction de l'exercice 8 – Vrai ou faux sur la distance à l'espérance

a) **Faux.** C'est vrai **en général**, pas toujours : un échantillon de 100 lancers peut, par hasard, tomber très près de μ, et un échantillon de 400 s'en écarter un peu plus. La simulation décrit une tendance, pas une certitude pour chaque échantillon.

b) **Vrai.** La distance typique se comporte comme $\frac{\sigma}{\sqrt{n}}$, et $\frac{\sigma}{\sqrt{9n}} = \frac{\sigma}{3\sqrt{n}}$.

c) **Vrai.** Un écart type nul signifie que la variable prend une seule valeur, égale à μ, avec la probabilité 1 : toutes les valeurs de l'échantillon valent μ, donc leur moyenne aussi.

d) **Faux.** Elle est proportionnelle à l'écart type σ de la variable : à taille égale, une variable plus dispersée donne des moyennes plus dispersées.

Réponse. a) faux ; b) vrai ; c) vrai ; d) faux.

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite.
Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).

Correction de l’exercice 9 – Estimer l’écart type à partir d’un échantillon

a) La somme des valeurs vaut 33, donc $m = 3{,}3$. La somme des carrés vaut $25 + 1 + 16 + 16 + 25 + 1 + 4 + 16 + 16 + 9 = 129$, d’où une moyenne des carrés de 12,9.

b) On applique la formule « moyenne des carrés moins carré de la moyenne » :

$$s = \sqrt{12{,}9 - 3{,}3^2} = \sqrt{12{,}9 - 10{,}89} = \sqrt{2{,}01} \approx 1{,}418$$

L’écart type de l’échantillon est une estimation de σ ; avec 10 valeurs seulement, il s’en écarte sensiblement (1,418 contre 1,708). Sur un grand échantillon, il en serait, sauf exception, beaucoup plus proche.

c) On suit la même formule :

```
from math import sqrt

def ecart_type(liste):
    n = len(liste)
    m = sum(liste) / n
    m2 = sum([x * x for x in liste]) / n
    return sqrt(m2 - m * m)
```

Réponse. a) $m = 3{,}3$, moyenne des carrés 12,9 ; b) $s = \sqrt{2{,}01} \approx 1{,}418$; c) *sqrt(m2 - m * m)*.

Correction de l’exercice 10 – Changer d’unité

a) Par linéarité, $E(Y) = 60\,E(X) = 45$ minutes. De plus $V(Y) = 60^2\,V(X)$, car $(Y - E(Y))^2 = 60^2 (X - \mu)^2$, donc $\sigma(Y) = 60 \times 0{,}2 = 12$ minutes.

b) Pour X, on obtient $\dfrac{0{,}2}{4} = 0{,}05$ heure ; pour Y,

$$\frac{12}{\sqrt{16}} = 3 \text{ minutes}$$

c) Les deux résultats décrivent la même chose : 0,05 heure, c’est 3 minutes. Changer d’unité multiplie l’espérance, l’écart type et la distance typique par le même facteur ; la précision de l’estimation ne dépend pas de l’unité choisie.

Réponse. a) $E(Y) = 45$, $\sigma(Y) = 12$; b) 0,05 h et 3 min ; c) c’est la même distance, exprimée dans deux unités.

Correction de l’exercice 11 – Moyenne de plusieurs échantillons

a) La somme des cinq moyennes vaut 17,45, d’où une moyenne de $\dfrac{17{,}45}{5} = 3{,}49$. Les cinq échantillons ayant la même taille, c’est la moyenne des $5 \times 200 = 1\,000$ lancers réunis.

b) Pour un échantillon de 200 lancers, $\dfrac{1{,}708}{\sqrt{200}} \approx 0{,}121$; pour l’échantillon regroupé,

$$\frac{1{,}708}{\sqrt{1\,000}} \approx 0{,}054$$

c) La moyenne regroupée s’écarte de 0,01 seulement de 3,5, bien moins que la distance typique 0,054 : elle est très proche de l’espérance. Les moyennes individuelles s’en écartaient jusqu’à 0,12, ce qui est de l’ordre de la distance typique pour 200 lancers.

Réponse. a) 3,49, moyenne de 1 000 lancers ; b) $\approx 0{,}121$ et $\approx 0{,}054$; c) oui, écart de 0,01.

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite.
Tout contrevenant s’expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).

Correction de l’exercice 12 – Lire un programme de simulation complet

a) *X()* renvoie 5 avec la probabilité 0,2 et 0 avec la probabilité 0,8. Donc $\mu = 5 \times 0{,}2 = 1$, $E(X^2) = 25 \times 0{,}2 = 5$, $V = 5 - 1 = 4$ et $\sigma = 2$.

b) L contient 200 nombres : les moyennes de 200 échantillons de taille 100. Au total, $200 \times 100 = 20\,000$ valeurs de X sont simulées.

c) Le premier affichage est la moyenne des 200 moyennes, c’est-à-dire la moyenne de 20 000 valeurs : elle doit être très proche de $\mu = 1$. Les moyennes ont un écart type d’environ

$$\frac{\sigma}{\sqrt{n}} = \frac{2}{10} = 0{,}2$$

Sur 200 échantillons, les plus extrêmes s’écartent en général de deux à trois écarts types : on peut s’attendre à un minimum vers 0,4 à 0,6 et à un maximum vers 1,4 à 1,6.

Réponse. a) 5 avec 0,2, 0 sinon ; $\mu = 1$, $\sigma = 2$; b) 200 moyennes, 20 000 valeurs ; c) environ 1 ; extrêmes à 2 ou 3 fois 0,2 de 1.

Correction de l’exercice 13 – Le temps d’attente au guichet

a) $E(T) = 0 \times 0{,}4 + 5 \times 0{,}3 + 10 \times 0{,}2 + 20 \times 0{,}1 = 1{,}5 + 2 + 2 = 5{,}5$ minutes. Puis $E(T^2) = 25 \times 0{,}3 + 100 \times 0{,}2 + 400 \times 0{,}1 = 7{,}5 + 20 + 40 = 67{,}5$, d’où

$$V(T) = 67{,}5 - 5{,}5^2 = 37{,}25 \qquad \sigma(T) \approx 6{,}10$$

b) La distance vaut $|6{,}9 - 5{,}5| = 1{,}4$ minute, et $\dfrac{6{,}10}{6} \approx 1{,}02$. La moyenne du jour s’écarte donc d’environ 1,4 écart type de l’espérance.

c) Non. Un écart de 1,4 écart type est courant : les simulations montrent que les moyennes d’échantillons s’écartent souvent d’un ou deux écarts types $\frac{\sigma}{\sqrt{n}}$. Avec seulement 36 clients et un temps d’attente très dispersé, ce jour-là n’a rien d’exceptionnel ; il faudrait observer plus de clients, ou plusieurs jours, pour conclure.

Réponse. a) $E(T) = 5{,}5$, $\sigma \approx 6{,}10$; b) distance 1,4, environ 1,4 fois $\frac{\sigma}{6} \approx 1{,}02$; c) non, écart ordinaire.

Correction de l’exercice 14 – Estimer une proportion inconnue

a) $\sigma = \sqrt{p(1-p)}$. La fonction $p \mapsto p(1-p) = -p^2 + p$ est un polynôme du second degré de coefficient dominant négatif, maximal en $p = \dfrac{0+1}{2} = \dfrac{1}{2}$ (milieu de ses racines 0 et 1), où il vaut $\frac{1}{4}$. Donc σ est maximal pour $p = \frac{1}{2}$, et vaut alors $\sqrt{\frac{1}{4}} = \frac{1}{2}$.

b) Pour tout $p \in [0\,;1]$, $\sigma \leq \frac{1}{2}$, donc

$$\frac{\sigma}{\sqrt{n}} \leq \frac{1}{2\sqrt{n}}$$

c) Il suffit que $\frac{1}{2\sqrt{n}} \leq 0{,}01$, soit $\sqrt{n} \geq 50$, c’est-à-dire $n \geq 2\,500$. Cette taille convient quelle que soit la proportion inconnue p.

Réponse. a) $\sigma = \sqrt{p(1-p)}$, maximal en $p = \frac{1}{2}$ où il vaut $\frac{1}{2}$; b) $\frac{\sigma}{\sqrt{n}} \leq \frac{1}{2\sqrt{n}}$; c) $n \geq 2\,500$.

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite.
Tout contrevenant s’expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).

Correction de l’exercice 15 – Synthèse : la masse des pommes

a) $E(M) = 150 \times 0{,}2 + 160 \times 0{,}3 + 170 \times 0{,}3 + 180 \times 0{,}2 = 30 + 48 + 51 + 36 = 165$ g. Les écarts à 165 valent -15, -5, 5, 15, d’où

$$V(M) = 0{,}2 \times 225 + 0{,}3 \times 25 + 0{,}3 \times 25 + 0{,}2 \times 225 = 105$$

et $\sigma(M) = \sqrt{105} \approx 10{,}25$ g.

b) Les 25 pommes d’un sachet forment un échantillon de taille 25 de M, et leur masse moyenne est la moyenne de cet échantillon. Sa distance typique à 165 g est d’environ $\dfrac{10{,}25}{5} = 2{,}05$ g.

c) On découpe un tirage *random()* selon les probabilités cumulées 0,2 ; 0,5 ; 0,8 :

```
from random import random

def pomme():
    u = random()
    if u < 0.2:
        return 150
    elif u < 0.5:
        return 160
    elif u < 0.8:
        return 170
    return 180

def sachet():
    return sum([pomme() for i in range(25)])
```

d) La masse totale attendue est $25 \times 165 = 4\,125$ g, et sa distance typique à cette valeur est 25 fois celle de la moyenne : $25 \times 2{,}05 \approx 51$ g.

e) Un sachet pèse en général 4 125 g à une cinquantaine de grammes près ; descendre sous 4 000 g demanderait un écart de 125 g, soit environ 2,4 fois la distance typique, ce qui est rare. L’affichage « 4 kg » est raisonnable, et légèrement prudent.

Réponse. a) $E(M) = 165$ g, $\sigma \approx 10{,}25$ g ; b) moyenne d’un échantillon de taille 25, distance typique $\approx 2{,}05$ g ; c) *sum([pomme() for i in range(25)])* ; d) 4 125 g, à environ 51 g près ; e) oui.

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite.
Tout contrevenant s’expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).