CORRIGÉ

Exercice 7 : *Consommation de chocolat et prix Nobel*

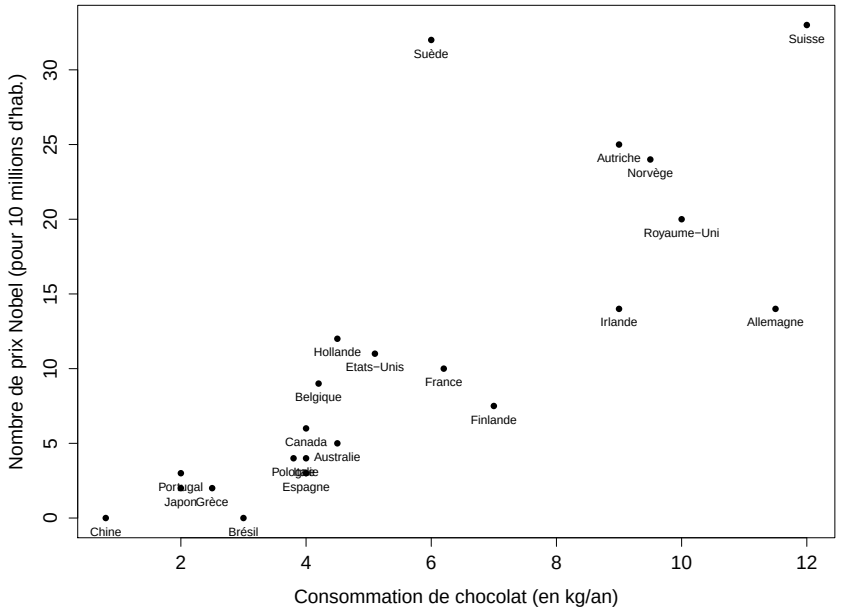
TABLE 1 – QUANTITÉ DE CHOCOLAT CONSOMMÉE (EN KILOS PAR PERSONNE/AN) ET NOMBRE DE PRIX NOBEL (POUR 10 MILLIONS D'HABITANTS) DANS 20 PAYS.

Pays	*Nobel*	*Cons. Choco*	*Pays*	*Nobel*	*Cons. Choco*
Allemagne	14	11,5	Finlande	7,5	7
Suisse	33	12	Suède	32	6
Belgique	9	4,2	Hollande	12	4,5
Etats-Unis	11	5,1	Italie	4	4
Canada	6	4	Royaume-Uni	20	10
Japon	2	2	Chine	0	0,8
France	10	6,2	Espagne	3	4
Brésil	0	3	Grèce	2	2,5
Portugal	3	2	Australie	5	4,5
Pologne	4	3,8	Autriche	25	9
Irlande	14	9	Norvège	24	9,5

Source : Messerli, F. (2012).

1. **Représenter le nuage de points (*X* : quantité de chocolat consommée, et *Y* : nombre de prix Nobel). Commenter.**

FIGURE 1 – NUAGE DE POINTS : QUANTITÉ DE CHOCOLAT CONSOMMÉE PAR HABITANT/AN ET NOMBRE DE PRIX NOBEL OBTENUS POUR 10 MILLIONS D'HABITANTS.



2. **Calculer la moyenne et la variance de la consommation de chocolat.**
   Il s'agit de la moyenne marginale correspondant à la moyenne de la distribution marginale de x. Toutefois, les données sont ici non-pondérées, autrement dit : $\bar{x} = \frac{1}{n}\sum_{i=1}^{p} x_i$, x_i indiquant la valeur prise par la variable x pour l'individu "i" (c'est-à-dire le pays "i").

Ainsi, dans notre cas :
$\overline{x} = \frac{1}{n}(x_1 + x_2 + ... + x_{22}) = \frac{11,5+12+...+9,5}{22} = 5,66$
En moyenne, les habitants des 22 pays étudiés consomment 5,66kg de chocolat par an.

De même pour la variance :
$V(x) = \frac{1}{n}\sum_{i=1}^{p}(x_i - \overline{x})^2$

Ainsi :
$V(x) = \frac{[(11,5-5,66)^2+(12-5,66)^2+...+(9,5-5,66)^2]}{22} = 10,32$
On a alors :
$\sigma_x = \sqrt{10,32} = 3,21$
En moyenne, l'écart entre la quantité de chocolat consommée annuellement par pays et la quantité moyenne consommée dans l'échantillon dans ces 22 pays était de 3,21kg par habitant.

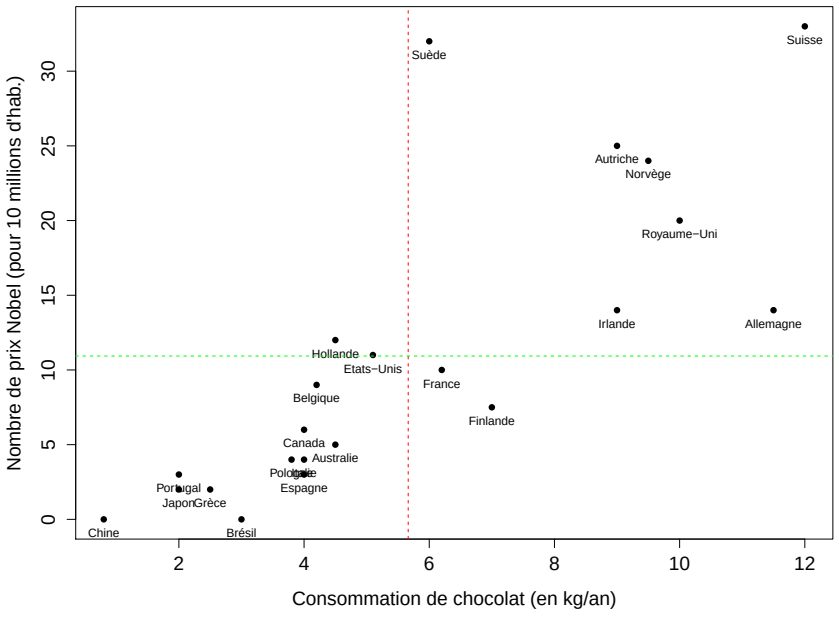
3. **Calculer la moyenne et la variance du nombre de prix Nobel pour 10 millions d'habitants.**
Pour les mêmes raisons qu'à la question précédente :
$\overline{y} = \frac{1}{n}(y_1 + y_2 + ... + y_{22}) = \frac{14+33+...+24}{22} = 10,93$
En moyenne, 10,93 prix Nobel sont obtenus pour 10 millions d'habitants au sein des 22 pays étudiés.

De même pour la variance :
$V(y) = \frac{1}{n}\sum_{j=1}^{q}(y_j - \overline{y})^2$

Ainsi :
$V(y) = \frac{[(14-10,93)^2+(33-10,93)^2+...+(24-10,93)^2]}{22} = 99,91$
On a alors :
$\sigma_y = \sqrt{99,91} = 9,996$
En moyenne, l'écart entre le nombre de prix Nobel obtenus pour 10 millions d'habitants par chaque pays et le nombre moyen de prix Nobel obtenus était de 10 pour 10 millions d'habitants.

4. **Placer le point moyen sur le nuage de points.**

FIGURE 2 – NUAGE DE POINTS ET POINT MOYEN : QUANTITÉ DE CHOCOLAT CONSOMMÉE PAR HABITANT/AN ET NOMBRE DE PRIX NOBEL OBTENUS POUR 10 MILLIONS D'HABITANTS.



Le point moyen correspond au centre de gravité du nuage de points, c'est-à-dire le point dont les coordonnées sont les moyennes des distributions marginales, soit : $(\overline{x};\overline{y})$
Dans notre cas, si P est notre point moyen : $P = (5,66;10,93)$

5. **Calculer la covariance et le coefficient de corrélation de Pearson. Qu'indiquent ces valeurs sur la relation entre les deux variables ?**
   — La notion de **covariance** est homogène à la notion de variance dans le cas d'une série à une dimension. Autrement dit, la covariance est la variation conjointe de deux variables.

   Ainsi, dans le cas de variables ponctuelles :

$$COV(X,Y) = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} x_i y_j - \overline{xy}$$

$$COV(X,Y) = \frac{1}{22}[(14.11,5) + (33.12) + ... + (24.9,5)] - 5,66.10,32 = 86,14 - 58,41 = 27,73$$

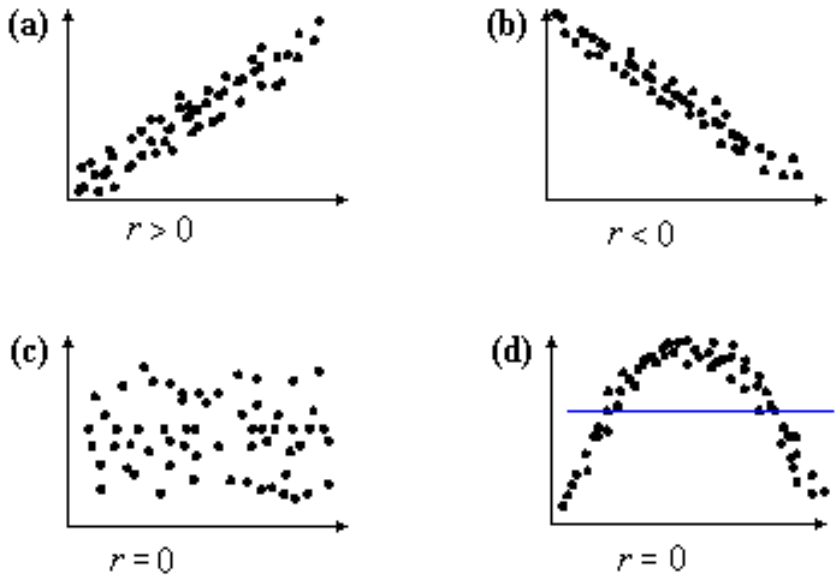
   La covariance est donc positive, les deux variables évoluent donc dans le même sens (lorsqu'une augmente l'autre aussi et inversement).

   Attention : la covariance est un outil imparfait, elle ne renseigne que sur le sens de la relation entre les deux variables, mais ne dit rien de l'intensité de la relation.

   — Le **coefficient de corrélation linéaire** (de Pearson) sert justement à mesurer cette notion d'intensité de la relation. Lorsque r est positif : les variables évoluent dans le même sens et inversement lorsque r est négatif. On dira qu'il existe une forte corrélation linéaire si r est proche de 1 ou -1 et inversement lorsque r se situe proche de 0.

   Attention : un coefficient de corrélation linéaire nul n'implique pas l'indépendance, mais l'indépendance implique une coefficient de corrélation nul, (voir graphiques ci-dessous).

FIGURE 3 – COEFFICIENTS DE CORRÉLATION LINÉAIRE, RELATIONS LINÉAIRES ET NON-LINÉAIRES.



Le coeefficient de corrélation linéaire de Pearson peut s'écrire de la façon suivante :

$$r = \sqrt{\frac{COV(X,Y)^2}{V(x)V(y)}} = \frac{COV(X,Y)}{\sigma_x \sigma_y}$$

Avec : $r \in [-1;1]$

Dans notre cas :

$$r = \frac{27,73}{3,21.9,996} = \frac{27,73}{32,09} = 0,86$$

A partir de quel seuil pouvons-nous dire que la corrélation est importante ? Pour répondre à cette question, nous devons démontrer à l'aide du rapport d'amélioration (A) que pour que r soit significatif d'un réel lien de dépendance, il doit être supérieur à 0,76.

En effet, $A = 1 - \sqrt{1-r^2}$ avec $\mid A \mid > 0,50$

On a : $A = 1 - \sqrt{1-0,76} = 0,51$

Ainsi, il existe une corrélation linéaire positive et très intense entre la consommation de chocolat an/hab. et le nombre de prix de Nobel pour 10 millions d'habitants dans les pays étudiés. Autrement dit, plus la consommation de chocolat augmente et plus le nombre de prix Nobel augmente et réciproquement.

6. **Déterminer les paramètres de la droite de régression ($D : y = \alpha x + \beta$, avec α le coefficient directeur et β l'ordonnée à l'origine de la droite de régression D).**
Il s'agit de déterminer une droite D pour laquelle la distance entre chaque point du nuage et la droite D soit minimale. Cette droite D, "résumant" l'information contenu dans le nuage de points, s'appelle alors la droite de régression.
L'ajustement linéaire consiste à minimiser les carrés des distances entre chaque point, de coordonnées (x ; y) et la droite de régression Autrement dit, nous utilisons la méthode des MCO. Les distances d_j sont prises parallèlement à l'axe des ordonnées.

En définitive, la droite D est telle que la somme des carrés des distances, d_j^2 soit minimale :

$$D : \sum_{j=1}^{q} d_j^2 min$$

$$D : \sum_{j=1}^{q} (y_j - \hat{y}_j)^2 min$$

Avec : $\hat{y_j}$ l'ordonnée d'un point quelconque M appartenant à la droite D.

Sachant que la droite D passe par le point moyen (fiche D2), nous connaissons désormais les coordonnées d'un point de cette droite : $(\overline{x};\overline{y})$.

Nous pouvons définir les paramètres de la droite D de la façon suivante :

$$\overline{y} = \alpha\overline{x} + \beta$$

Et :

$$\beta = \overline{y} - \alpha\overline{x}$$

Ainsi, α définit la pente de la droite D, et β constitue l'ordonnée à l'origine.

Or, nous savons d'après la fiche (D3) que :

$$\alpha = \frac{COV(X,Y)}{V(x)}$$

Et que :

$$\beta = \overline{y} - \alpha\overline{x}$$

Ainsi, pour la droite D, nous avons :

$$y = \frac{COV(X,Y)}{V(x)}x + \overline{y} - \overline{x}\frac{COV(X,Y)}{V(x)}$$

$$y = \overline{y} + \frac{COV(X,Y)}{V(x)}.(x - \overline{x})$$

Dans notre cas, pour la pente de la droite D :

$$\alpha = \frac{27,73}{10,32} = 2,69$$

Pour l'ordonnée à l'origine de la droite D :

$$\beta = 10,93 - (2,69.5,66) = -4,30$$

Ainsi :

$$D : y = 2,69x - 4,30$$

Nous pouvons alors estimer qu'en moyenne, pour un kilo supplémentaire de chocolat consommé par an/hab., le nombre de prix Nobel pour 10 millions d'hab. augmentera en moyenne de 2,69.

7. **Démontrer que :**

$$\alpha = r\frac{\sigma_y}{\sigma_x}$$

***Avec :** α **le coefficient directeur de la droite de régression D,** r **le coefficient de corrélation de Pearson,** σ_x **et** σ_y **respectivement la racine carrée des variances de la variable** X **et de** Y**.**

Nous savons désormais (*conf.* fiche *D*3) que :

$$\alpha = \frac{COV(X,Y)}{V(x)}$$

Et que :

$$r = \frac{COV(X,Y)}{\sigma_x\sigma_y}$$

Et : $V(x) = \sigma_x^2$

On peut donc écrire :

$$\alpha = \frac{COV(X,Y)}{\sigma_x^2} = \frac{COV(X,Y)}{\sigma_x\sigma_x}$$

Donc :

$$\sigma_x\alpha = \frac{COV(X,Y)}{\sigma_x}$$

Or :

$$\sigma_y r = \frac{COV(X,Y)}{\sigma_x}$$

Finallement :

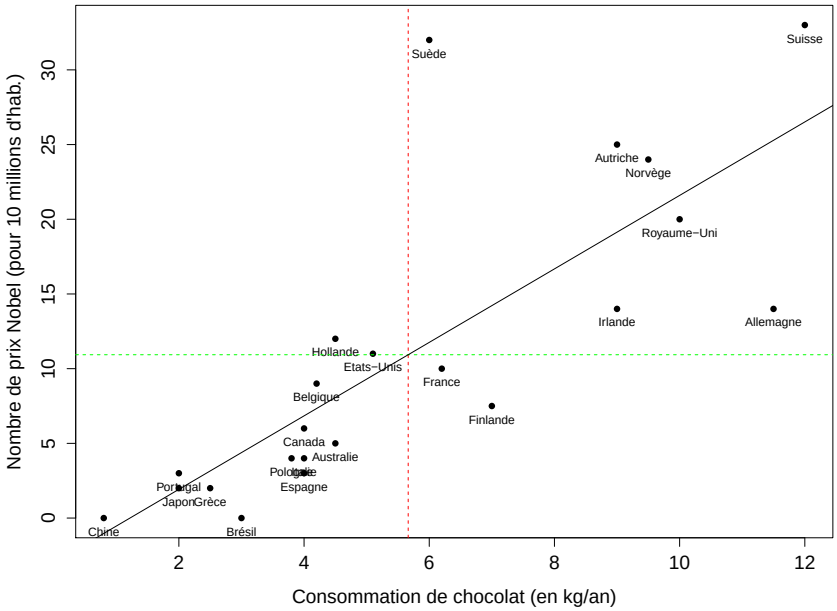
$$\alpha\sigma_x = r\sigma_y = \frac{COV(X,Y)}{\sigma_x}$$

Donc :

$$\alpha = r\frac{\sigma_y}{\sigma_x}$$

8. **Représenter cette droite de régression au sein du nuage tracé à la question 1.**

FIGURE 4 – NUAGE DE POINTS, POINT MOYEN ET DROITE DE RÉGRESSION : QUANTITÉ DE CHOCOLAT CONSOMMÉE PAR HABITANT/AN ET NOMBRE DE PRIX NOBEL OBTENUS POUR 10 MILLIONS D'HABITANTS.



9. **Estimer le nombre de prix Nobel pour 10 millions d'habitants d'un pays dans lequel la consommation moyenne de chocolat serait de 8 kilogrammes par personne et par an.**
   Nous connaissons la relation linéaire existante entre la variable X et Y : $D : \hat{y} = 2,69x - 4,30$.
   Or, nous cherchons à prédire une valeur de y (soit $\hat{y}$) correspondante à la valeur $x = 8$.

   Ainsi, pour $x = 8$:

$$\hat{y} = 2,69.8 - 4,30$$

$$\hat{y} = 17,22$$

Références :

Messerli, F. (2012) Chocolate Consumption, Cognitive Function, and Nobel Laureates. *The New England Journal of Medecine* 367 :1562-1564.