# CORRIGÉ

## Exercice 8 : *Corrélation et causalité*

1. **Interpréter précisément le $r = 0,791$ sur la figure 1.**
   Il s'agit du coefficient de corrélation de Pearson. Ce dernier indique l'intensité et le sens de la relation linéaire entre les variables $X$ (consommation de chocolat en kg/habitant et par an) et $Y$ (nombre de prix Nobel pour 10 millions d'habitants).

   Nous savons également que :

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

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

   Le coefficient de corrélation est de $0,791$, autrement dit la relation linéaire est forte et indique donc que plus la consommation de chocolat kg/hab. par an augmente et plus le nombre de prix Nobel pour 10 millions d'habitants augmente (et réciproquement).

2. **D'après les résultats obtenus à l'exercice 3 et la figure 1, peut-on dire que la consommation de chocolat est la cause de l'obtention de prix Nobel ? Vous expliquerez votre raisonnement en vous aidant des résultats de Messerli (2012).**
   L'exercice nous pousse à penser que la consommation de chocolat des habitants d'un pays serait la cause du nombre de prix Nobel obtenus. Ainsi, une recommandation de politique publique pourrait être : consommer du chocolat !

   Toutefois, cette analyse est **complètement erronée**. Une corrélation n'implique pas nécessairement la causalité. Lorsque nous observons une corrélation, indépendamment de son intensité, il est impossible de conclure quant à la causalité d'une variable sur l'autre. Plusieurs raisons expliquent cette limite :

   (a) Tout d'abord le problème de la **simultanéité** ($X$ donne $Y$ puis $Y$ va donner $X$ et ainsi de suite...) qui s'exprime sous la forme d'une boucle de rétroaction. La corrélation entre deux variables $X$ et $Y$ signifie littéralement que la variation de la première est en relation avec la variation de la seconde (dans le même sens, ou en sens inverse), sans qu'aucune conclusion puisse être tirée sur une quelconque causalité directe de $X$ sur $Y$ ou inversement. Dans le cas qui nous occupe ici, par exemple, l'augmentation de la consommation de chocolat peut être la cause de l'augmentation du nombre de prix Nobel (le chocolat ayant des effets cognitifs positifs) mais l'augmentation du nombre de prix Nobel peut également induire une plus forte consommation en raison même de ces effets recherchés.

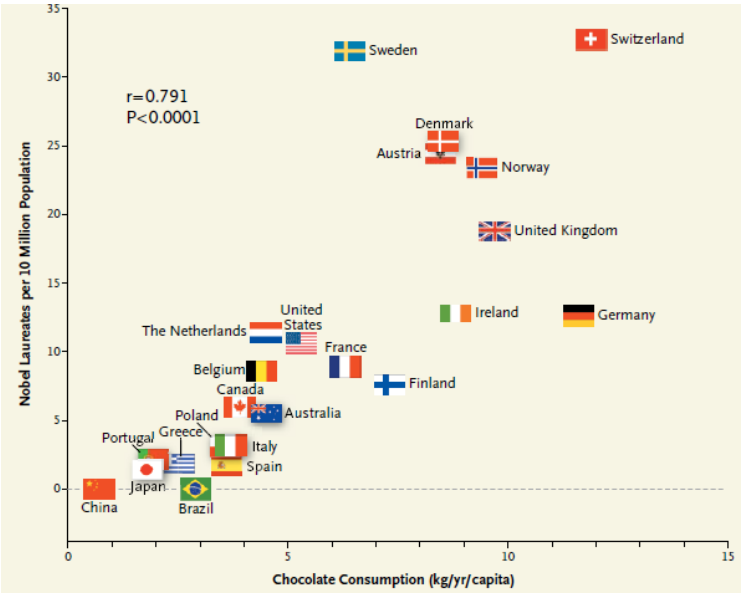
   (b) Nous pouvons tout aussi bien penser que $X$ est la cause de $Y$ tout comme $Y$ peut être la cause de $X$, il s'agit alors du problème de **causalité inverse**. Par exemple, la corrélation entre la consommation de chocolat ($X$) et le poids ($Y$) vue à l'exercice précédent peut s'entendre de deux façons. Nous pouvons imaginer que si les individus consomment de grandes quantités de chocolat ils auront tendance à grossir. A l'inverse, les individus ayant un poids plus élevé vont avoir tendance à manger plus (leur organisme nécessitant plus de calories) et donc auront tendance à ingurgiter des quantités plus importantes de chocolat, lorsqu'ils en mangent.

   (c) Un autre problème réside dans les **chaînes de causalités**. Les variables $X$ et $Y$ peuvent alors avoir effectivement une relation causale mais indirecte, autrement dit séparée par un ensemble de relations avec des variables intermédiaires. Une autre raison, ces mêmes pays dégagent des ressources fiscales importantes pour financer l'enseignement supérieur et la recherche publique et privée, ce qui va produire par conséquent un nombre important de prix Nobel.

(d) Un autre problème réside de la présence d'une cause commune (**variable omise**) expliquant à la fois la variable $X$ et la variable $Y$. Dans le cas étudié, les prix Nobel remis aux chercheurs des pays en développement sont extrêmement rares, il y a donc un déterminisme géographique à la source de la variable $Y$. Par ailleurs, le chocolat est un produit historiquement capté par les espagnols en Amérique latine et démocratisé dans les différents pays d'Europe à l'époque industrielle (les plus gros consommateurs mondiaux sont : la Suisse, l'Allemagne, l'Irlande, le Royaume-Uni, la Norvège...). Sa consommation est donc liée aux pays ayant connu un développement industriel basé ou en lien avec un modèle impérialiste, captant les ressources des pays en développement (colonisés), dont le chocolat fait partie intégrante. Au regard des bénéfices liés à cette histoire coloniale et bien que ce mode de captation systématique ait disparu, les ex–pays colonisateurs apparaissent comme ceux ayant un niveau de richesse très élevé (RNB/hab. et IDH par exemple) pouvant expliquer la part consacré à la recherche publique et privée dans ces États. Ce déterminisme géographique est donc aussi à la source des variable $X$ et $Y$, il s'agit donc d'une dimension omise.

(e) Enfin, le **hasard** a pu créer cette corrélation statistique sans aucun fondement causal, ce que semble bien traduire finalement la relation observée entre la consommation de chocolat et le nombre de prix Nobel pour les pays observés. D'autres exemples sont disponibles sur le site Spurious Correlations, faisant état de ce problème de croisement hasardeux de distributions ayant des variations similaires sans qu'il n'existe un quelconque lien entre elles (en particulier : entre le nombre de personnes mortes noyées dans une piscine et le nombre de films dans lesquels Nicolas Cage apparait ! Les détracteurs de Nicolas Cage n'en sont pas au point de se noyer à chaque apparition de l'acteur !).

FIGURE 1 – CORRÉLATION ENTRE LA QUANTITÉ DE CHOCOLAT CONSOMMÉE PAR HABITANT/AN ET LE NOMBRE DE PRIX NOBEL OBTENUS POUR 10 MILLIONS D'HABITANTS.




*Source :* Messerli (2012).

Références :

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