ECE 2 - Année 2020-2021
Lycée français de Vienne
Mathématiques - F. Gaunard
`http://frederic.gaunard.com`

T.P. n°3

Statistiques descriptives bivariées

☞ ***Mots-clés**: covariance empirique, droite de régression*

1 - Introduction

Exemple 1 : émissions de CO_2 et surface de forêt dans le monde

On a récupéré sur le site *The World Bank* (et on a converti pour une utilisation sous `SciLab`) les données concernant l'évolution des émissions de CO_2 dans le monde (en kT) ainsi que l'évolution de la surface du monde couverte par de la forêt (en km^2) que l'on propose de retrouver ici.

(1) Combien la série statistique `X` comporte-t-elle de données ? Même question pour `Y`.
(2) Définir une série statistique `Z` de même longueur que `Y`, concernant la même période.
(3) Calculer les moyennes $\overline{y}$ et $\overline{z}$ de ces deux séries.
(4) Représenter le *nuage de points* de la série statistique `(Y,Z)`. Y ajouter le *point moyen*, de manière visible.
(5) Comment calculer la covariance empirique de chacune des séries statistiques précédentes?
(6) Calculer le coefficient de corrélation linéaire (empirique) du couple statistique `(Z,Y)`, que l'on notera `rhoYZ`. Interpréter.
(7) Représenter (en superposition) en rouge sur le graphique précédent la droite d'équation $y = ax + b$ où

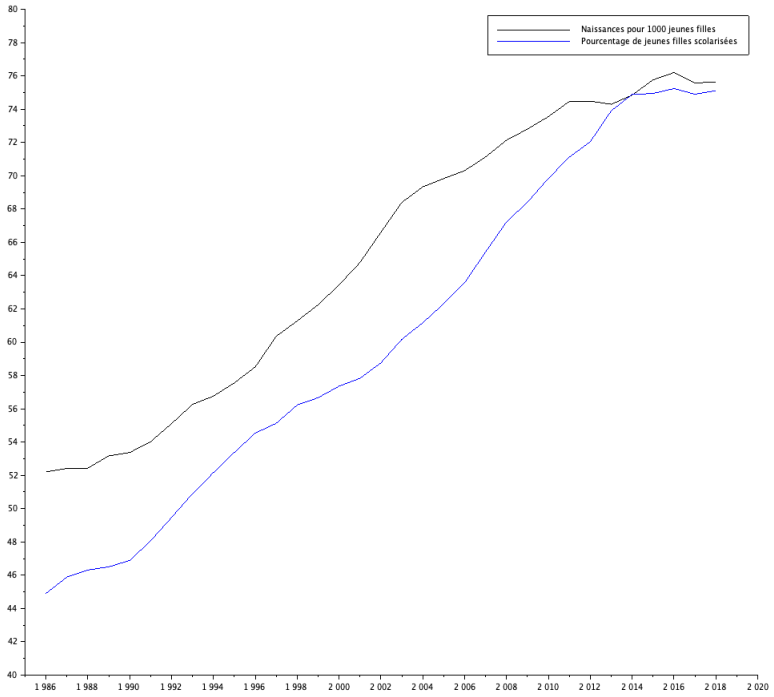
$$a = \frac{\texttt{rhoYZ}}{\texttt{corr(Y,Y,1)}}, \quad \text{et} \quad b = \overline{z} - a\overline{y}.$$

(8) Les données pour la surface de forêt en 2015 et 2016 sont disponibles sur le même site (resp. 39991324.6 et 39958245.9 km^2). Peut-on prévoir les émissions de $\texttt{CO}_2$ pour les mêmes années ?

Exemple 2 : taux de scolarisation et *fertilité* adolescente

Les courbes ci-contre représentent l'évolution, sur la période 1986-2018, du pourcentage de scolarisation des jeunes filles dans l'enseignement secondaire dans le monde ainsi que le nombre, pour 1000 jeunes filles entre 15 et 19 ans du nombre de naissances, dans le monde pour la même période.

Ces données proviennent du site *The World Bank* et sont disponibles pour une utilisation sous `SciLab` ici.



(1) Représenter le *nuage de points* correspondant à cette série statistique (X, Y)
(2) Calculer les variance(s) et covariance empiriques des éléments de la série statistique (X, Y) puis le coefficient de corrélation linéaire (empirique) du couple (X, Y).
(3) Déterminer et représenter graphiquement la droite de régression.

2 - Support et bla bla théorique

On appelle **nuage de points** associé à la série statistique (X, Y) l'ensemble des points M_k de coordonnées (x_k, y_k) (pour $1 \leq k \leq n$) tracés dans un repère orthonormé du plan (où $X = (x_k)$ et $Y = (y_k)$).

L'examen du nuage de points permet de faire des constatations qualitatives:

- est-il concentré ou dispersé?
- relève-t-on une tendance?
- y a-t-il des valeurs *a priori* aberrantes?

Le **point moyen** du nuage est le point de coordonnées $(\overline{x}, \overline{y})$, où $\overline{x}$ désigne la moyenne empirique des x_k et $\overline{y}$ celle des y_k:

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k.$$

☞ On utilise la fonction `plot2d()` pour représenter le nuage de points ainsi que le point moyen (que l'on différencie).

☞ Les variances et covariance empiriques du couple sont définies par

$$\sigma_X^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \overline{x})^2, \qquad \mathrm{cov}(X, Y) = \frac{1}{n} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y}).$$

☞ La commande `corr(x,y,1)` renvoie également la covariance du couple `(x,y)`. En particulier, `corr(x,x,1)` permet d'obtenir la variance de `x`.

On se place dans la situation où l'on souhaite savoir si la la "courbe" de Y en fonction de X peut être *approximée* par une droite. On cherche donc deux constante a et b telles que

$$Y = aX + b + \varepsilon.$$

Théorème 1. *Soit ρ le coefficient de corrélation linéaire du couple (x, y). Alors*
(i) $\rho \in [-1; 1]$;
(ii) $\rho = \pm 1$ si et seulement si la régression $Y = aX + b$ est exacte.

☞ Par conséquent, si $|\rho|$ est proche de 1 **et qu'on a visualisé une relation linéaire entre les données**, on peut confirmer qu'il y a bien corrélation linéaire entre X et Y.

☞ En sciences humaines et en sciences économiques, une valeur de $|\rho|$ de l'ordre de $0,85$ est souvent considérée comme bonne.

On utilise alors la *méthode des moindres carrés* qui nous donne l'équation de la droite la plus proche des points en terme de distance, c'est à dire l'unique droite D d'équation $y = ax + b$ qui rend minimale la somme des carrés des erreurs d'ajustement

$$d^2(a, b) = \frac{1}{n} \sum_{i=1}^{n} (y_i - (ax_i + b))^2 .$$

Le résultat suivant donne la valeur de a et b et est admis. Sa démonstration repose sur des outils d'*algèbre bilinéaire* non accessibles avec le programme d'ECE (la série statistique $aX + b\mathbf{1}$ est le *projeté orthogonal* de Y sur le plan de $\mathbb{R}^n$ engendré par X et $\mathbf{1}$ pour le produit scalaire $\langle X, Y\rangle = E(XY)$).

Théorème 2. *La droite la plus proche du nuage de points associé au couple (x, y) est la droite d'équation $y = ax + b$ avec*

$$a = \frac{\mathrm{cov}(x, y)}{V(x)}, \quad \text{et} \quad b = \overline{y} - a \times \overline{x}.$$

En particulier, cette droite passe par le point moyen $(\overline{x}, \overline{y})$.

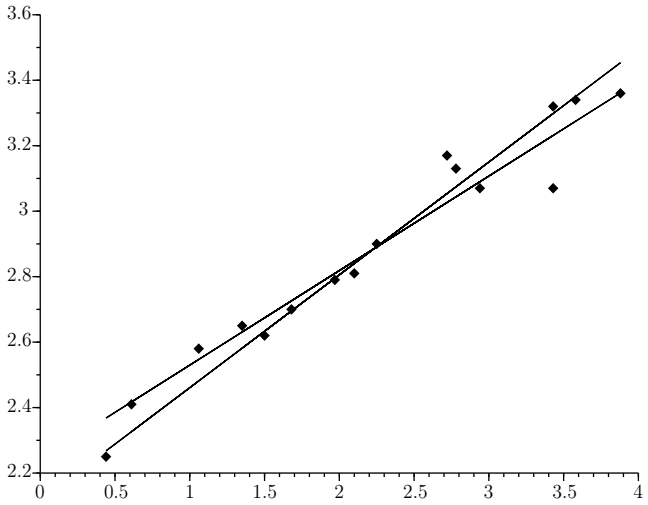
2 - Un exercice d'annale - HEC 2016

On rappelle qu'en `SciLab`, si $(v_i)_{1\leq i\leq n}$ et $(w_i)_{1\leq i\leq n}$ sont deux séries statistiques, alors la variance de $(v_i)_{1\leq i\leq n}$ est calculable par la commande `variance(v)` et la covariance de $(v_i, w_i)_{1\leq i\leq n}$ est calculable par `corr(v,w,1)`.

(1) On a relevé pour $n = 16$ entreprises qui produisent le bien considéré à l'époque donnée, les deux séries statistiques $(u_i)_{1\leq i\leq n}$ et $(t_i)_{1\leq i\leq n}$ reproduites dans les lignes (1) et (2) du code `SciLab` suivant dont la ligne (5) est incomplète :

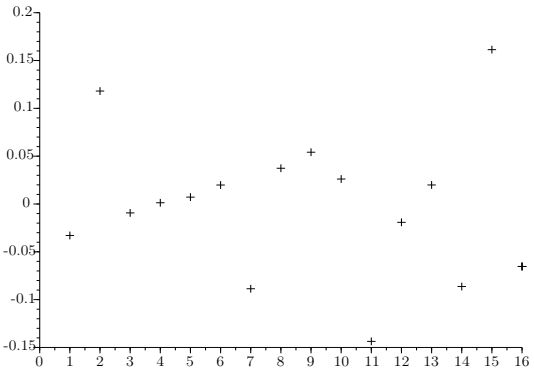
```
(1) u=[1.06,0.44,2.25,3.88,0.61,1.97,3.43,2.10,1.50,1.68,2.72,1.35,2.94,2.78,3.43
,3.58]
(2) t=[2.58,2.25,2.90,3.36,2.41,2.79,3.32,2.81,2.62,2.70,3.17,2.65,3.07,3.13,3.07
,3.34]
(3) plot2d(u,t,-4) // -4 signifie que les points sont représentés par des
losanges.
(4) plot2d(u,corr(u,t,1)/variance(u)*u+mean(t)-corr(u,t,1)/variance(u)*mean(u))
// équation de la droite de régression de t en u.
(5) plot2d(u,.....)
// équation de la  droite de régression de u en t.
```

Le code précédent complété par la ligne (5) donne alors la figure suivante :



(a) Compléter la ligne (5) du code permettant d'obtenir la figure précédente.
(b) Interpréter le point d'intersection des deux droites de régression.
(c) Estimer graphiquement les moyennes empiriques $\overline{u}$ et $\overline{t}$.
(d) Le coefficient de corrélation empirique de la série statistique double $(u_i, t_i)_{1\leq i\leq 16}$ est-il plus proche de -1, de 1 ou de 0 ?
(e) On reprend les lignes (1) et (2) du code précédent que l'on complète par les instructions (6) à (11) qui suivent et on obtient le graphique ci-dessous :

```
(6) a0=corr(u,t,1)/variance(u)
(7) b0=mean(t)-corr(u,t,1)/variance
(u)*mean(u)
(8) t0=a0*u+b0
(9) e=t0-t
(10) p=1:16
(11) plot2d(p,e,-1)
```



Que représente ce graphique ? Quelle valeur peut-on conjecturer pour la moyenne des ordonnées des 16 points obtenus sur le graphique ? Déterminer mathématiquement la valeur de cette moyenne.

3 - Un autre exemple : évolution du cours de titres du CAC 40

Un investisseur ayant nouvellement fait fortune se prend l'envie de *boursicoter*. Souhaitant naïvement réaliser une bonne diversification de son portefeuille boursier, celui-ci souhaite sélectionner des actifs qu'il estimera peu *corrélés*, espérant ainsi *réduire le risque* de son portefeuille.

Il récupère sur les sites dédiés (Boursorama, ABCBourse, BFM) les cours de clôture de différents titres: Société Générale (SG), AirFrance KLM (AFKLM) et BNP Paribas, pour chacun des 23 jours d'ouverture de la Bourse de Paris que comptait le mois d'octobre 2018, ici reproduits.

Date	AF KLM - cours de clôture (euros)	SG - cours de clôture (euros)	BNP - cours de clôture (euros)
01/10/18	8,614	36,595	51,99
02/10/18	8,43	36,54	51,98
03/10/18	8,584	36,755	52,14
04/10/18	8,64	37,00	52,30
05/10/18	8,442	36,81	51,85
08/10/18	8,394	36,295	51,15
09/10/18	8,83	36,25	51,23
10/10/18	8,482	36,65	51,47
11/10/18	7,994	36,07	50,19
12/10/18	8,224	35,66	49,915
15/10/18	7,956	35,34	49,50
16/10/18	7,982	35,41	49,72
17/10/18	8,09	35,145	49,60
18/10/18	8,478	34,895	49,225
19/10/18	8,224	34,45	48,725
22/10/18	8,146	34,09	48,305
23/10/18	8,022	33,60	47,56
24/10/18	8,20	32,705	46,46
25/10/18	8,366	32,885	47,155
26/10/18	8,366	32,37	46,88
29/10/18	8,404	32,63	47,475
30/10/18	8,492	32,11	46,145
31/10/18	8,56	32,48	46,13

Afin d'éviter au lecteur d'avoir à retaper toutes les données dans `SciLab`, celles-ci sont disponibles ici.

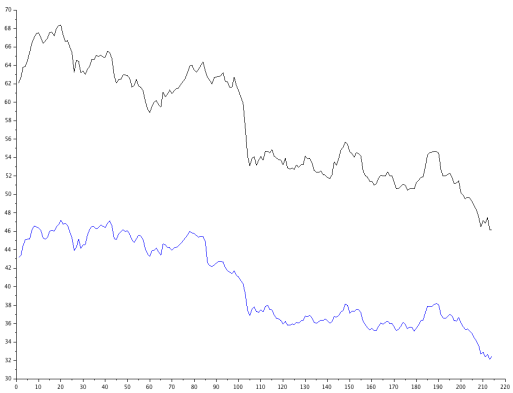
- La série statistique `X` correspond aux cours successifs du titre de *Air France KLM*;
- La série statistique `Y` correspond aux cours successifs du titre de *Société Générale*;
- La série statistique `Z` correspond aux cours successifs du titre de *BNP Paribas*.

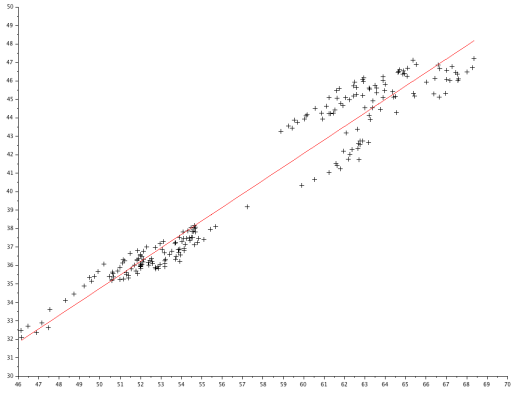
(1) Représenter sur un même graphique l'évolution de ces cours au mois d'octobre pour les titres des deux banques. Que semble-t-on observer?

(2) On cherche à mesurer la corrélation linéaire de Y et Z.
   (a) Représenter le nuage de points associé au couple de séries statistiques (Y, Z). *On représentera de manière différente le* ***point moyen*** *du nuage.* Commenter.
   (b) Calculer les variances empiriques de Y et Z, puis la covariance empirique puis enfin le coefficient de correlation linéaire empirique $\rho_{Y,Z}$. Commenter.
   (c) Déterminer l'équation de la droite de régression et la représenter en superposition du nuage de points.

(3) Faire la même étude comparative pour X et Y.

On ne résiste pas à l'envie de présenter les mêmes graphiques mais pour la période du 1er Janvier au 31 Octobre 2018.





On trouve $\rho_{Y,Z} = 0.9725222$.

4 - Un exercice de DS

Inspiré par **EDHEC 2019** et extrait de **CB n°3, sujet A**, Automne 2019.

Soit n un entier naturel supérieur ou égal à 3.
Une urne contient une boule noire non numérotée et $n-1$ boules blanches, dont $n-2$ portent le numéro 0 et une porte le numéro 1. On extrait ces boules au hasard, une à une, **sans remise**, jusqu'à l'apparition de la boule noire.

On note X la variable aléatoire égale au rang d'apparition de la boule noire et Y la variable aléatoire qui vaut 1 si la boule numérotée 1 a été piochée lors de l'expérience précédente, et qui vaut 0 sinon.

On rappelle qu'en `SciLab`, la commande `grand(1,1,'uin',a,b)` simule une variable aléatoire suivant la loi uniforme sur $[\![a,b]\!]$.

(1) Compléter la fonction `SciLab` suivant afin qu'il simule l'expérience aléatoire décrite dans cet exercice et renvoie les valeurs prises par les variables X et Y.
On admettra que la boule noire est codée tout au long de ce script par le nombre `nB+1`, où `nB` désigne le nombre de boules blanches.

```
function [x,y]=XY(n)
nB=n-1
x=1
y=.................
u=grand(1,1,'uin',1,nB+1)
while u<nB+1
        nB=...........
        if u==1 then y=..............
        end
        u=grand(1,1,'uin',1,........)
        x=.............
end
endfunction
```

(2) On génère un échantillon de taille 1000 du couple (X,Y). Compléter le script suivant afin de calculer la covariance empirique de l'échantillon obtenu.

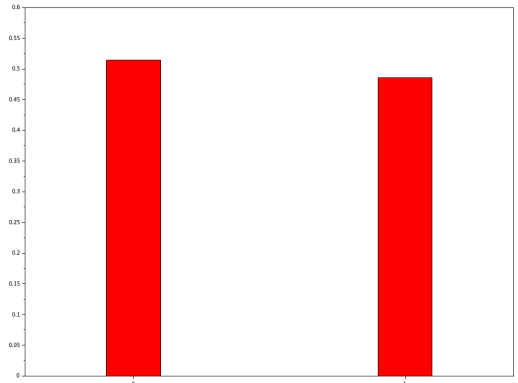
```
n=input('n=?')
X=zeros(1, 1000);
Y=zeros(1, 1000);
for k=1:1000
    [X(k), Y(k)]=XY(n)
end
covXY=mean(.............................)
disp(covXY)
```

(3) En rentrant $n=4$, le programme affiche `0.449475`. Que peut-on conjecturer?

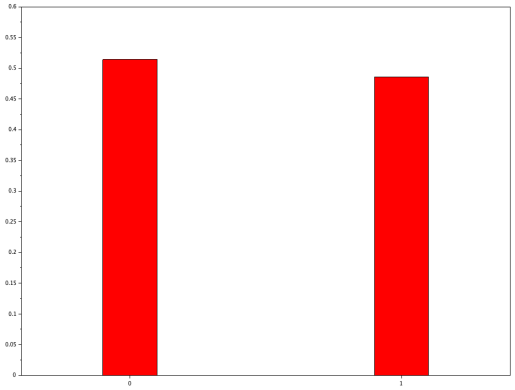
(4) On remplace les deux dernières lignes du programme précédent par le script suivant que l'on exécute successivement pour $n=4$, $n=5$, $n=10$ et $n=50$.

```
T=tabul(Y, 'i');
T(:,2)=T(:,2)/1000;
bar(T(:,1), T(:, 2), width=0.2, 'red')
```

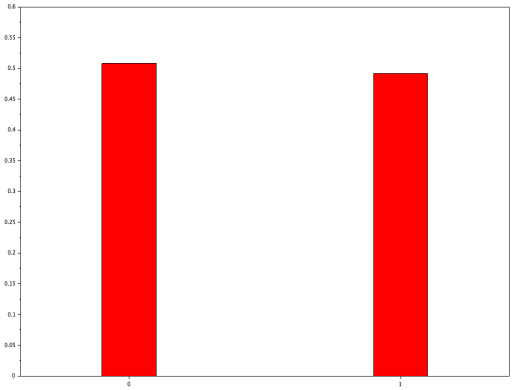
`SciLab` affiche alors respectivement les quatre figures ci-dessous



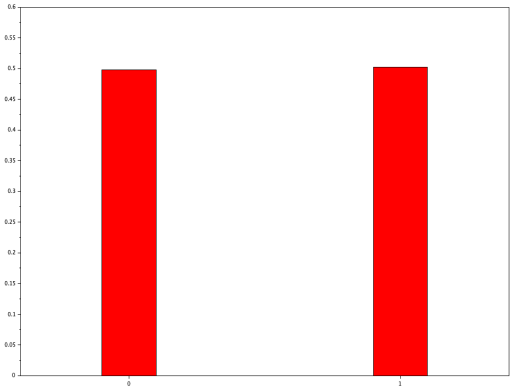
$n = 4$



$n = 5$



$n = 10$

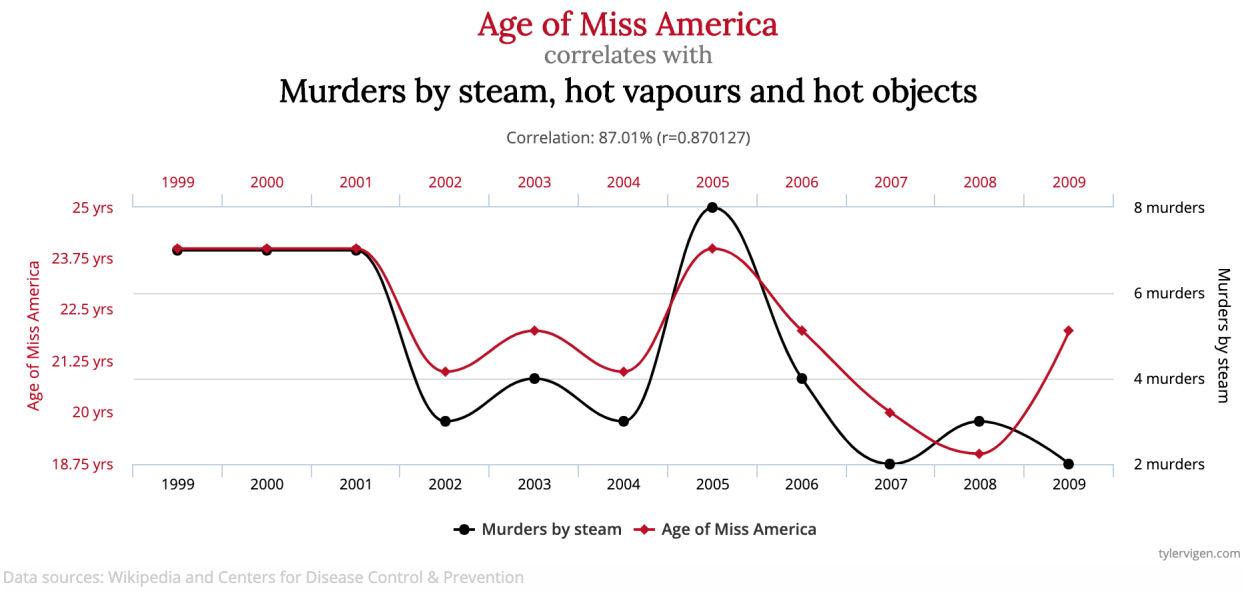


$n = 50$

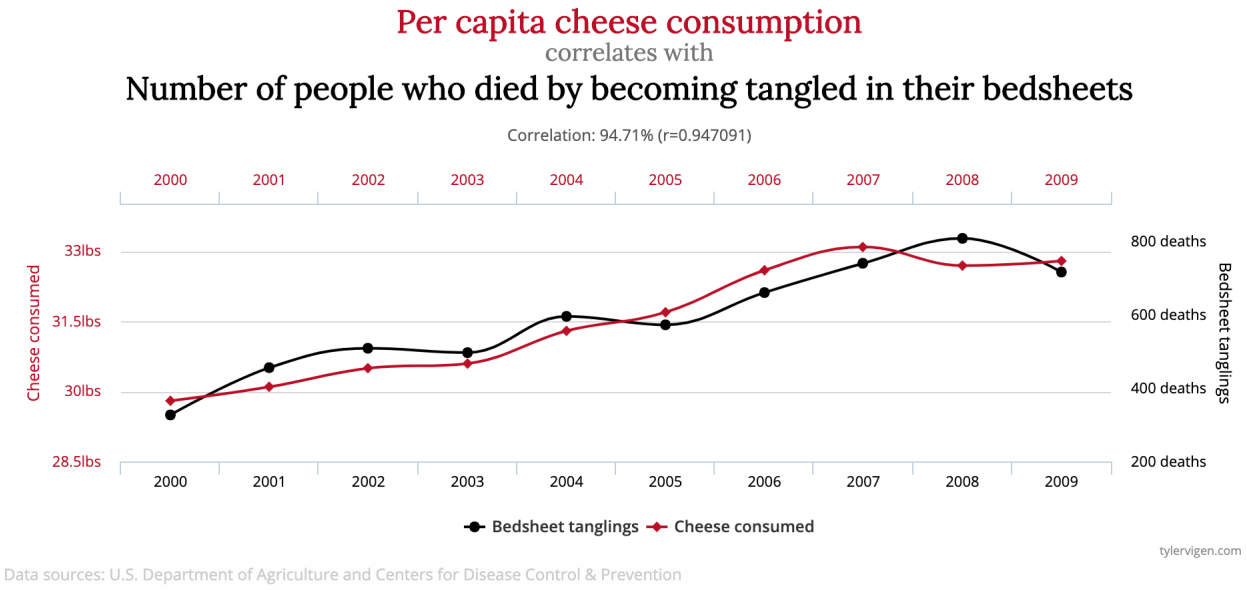
Que peut-on alors conjecturer quant à la la loi de Y?

5 - De la différence entre corrélation et causalité

L'objectif de ce cours est de présenter quelques outils du calcul statistique; nullement de développer des théories *abracadabrantesques*.
On évitera donc toute interprétation hasardeuse des résultats obtenus. Une *corrélation* est un lien statistique, sans qu'on se demande quelle variable agit sur l'autre. Une *causalité* est un lien qui affirme qu'une variable agit sur une autre. Il y a bien une corrélation entre la quantité de fromage consommée par habitant et le nombre de morts par étouffement entre les draps. Cependant, il semble difficile d'affirmer qu'il y a une causalité.
On peut par exemple présenter deux données qui n'ont ni la même échelle ni la même unité. En coupant les axes des ordonnées (à droite et à gauche), on peut superposer deux courbes qui n'ont rien à voir et laisser penser qu'elles ont une influence l'une sur l'autre.
Le site internet *Spurious correlations* propose quelques exemples rigolos pour mettre en lumière cette remarque.



Data sources: Wikipedia and Centers for Disease Control & Prevention



Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention