L1 - SOCIOLOGIE 2018-2019- CORRIGÉ DU TD2

COMMENTAIRE : le but de ces notes est d'expliquer sur les exercices de la feuille TD2 les techniques de calcul de l'analyse statistique en deux variables, en utilisant éventuellement les fonctions des calculatrices. Les résultats numériques que nous donnons sont fournis à titre indicatif, et ne dispensent pas les étudiants de faire eux-mêmes leurs propres calculs. D'ailleurs, nos résultats peuvent tout à fait contenir des erreurs.

Sur l'usage des calculatrices classiques pour les exercices de cette feuille de TD. *On rappelle d'abord comment utiliser les fonctions de listes des calculatrices CASIO et TI pour s'aider dans les exercices d'étude de corrélation de deux variables.*

a) On remplit en mode STATISTIQUES les listes L_1 et L_2 des valeurs respectives des variables X et Y, *en ayant au préalable vidé soigneusement L_1 et L_2 des résidus des exercices précédents !*

b) On *ordonne les deux listes par ordre croissant* : plus précisément, on dit que la liste de base est L_1, et la deuxième liste est L_2. Bien que ce ne soit pas indispensable pour le calcul du barycentre et des droites $D_{Y/X}$ et $D_{X/Y}$, ça l'est pour le calcul de la droite de Mayer, et utile pour la représentation graphique du nuage de points. *Il est donc bon d'avoir le réflexe de le faire dès le début.*

   (a) CASIO : Menu STAT→SRT-A→Répondre 2 (pour deux listes)→Répondre 1 (pour dire que la liste de base à ordonner est L_1)→Répondre 2 (pour la seconde liste L_2).

   (b) TI : STAT→EDIT→SortA(L1,L2).

c) Puis on procède à l'aide de la fonction CALC du menu STAT pour l'analyse *en deux variables* :

   (a) CASIO : Menu STAT→CALC puis *vérifier les settings* : SET→2Var XList : List 1, 2Var YList : List2, 2Var Freq : 1. Valider par EXE, puis 2Var en bas de l'écran avec les colonnes contenant L_1 et L_2.

   (b) TI : STAT→CALC→2-Var Stats (L_1,L_2).

   Ces fonctions donnent la *taille* n de l'échantillon (le nombre de points du nuage de points), les *moyennes* respectives $\bar{x} = m(X)$ et $\bar{y} = m(Y)$, la somme $\sum xy = \sum_{i=1}^{n} X_i Y_i$ (utile pour le calcul de la *covariance* $C(X,Y)$) et les *écarts-types* $\sigma(X)$ et $\sigma(Y)$ (qui permettent de connaître les *variances* $V(X) = \sigma(X)^2$ et $V(Y) = \sigma(Y)^2$). Donc le *centre de gravité* est le point $G = (\bar{x}, \bar{y}) = (m(X), m(Y))$.

   De même, la *covariance* est $C(X,Y) = m(XY) - m(X)\, m(Y) = \dfrac{1}{n}\sum xy - \bar{x} \cdot \bar{y}$.

d) Pour le calcul des *droites de régression $D_{Y/X}$ et $D_{X/Y}$, et du coefficient de corrélation linéaire* $r(X,Y)$, on peut procéder comme suit, d'abord pour la droite $D_{Y/X}$:

   (a) CASIO : Menu STAT→CALC (avec les même settings)→REG→X

   (b) TI : STAT→LinReg($ax+b$) L_1,L_2.
   On trouve les coefficients a et b de l'équation $Y = aX + b$ de la droite de régression $D_{Y/X}$, ainsi que le coefficient de corrélation linéaire $r(X,Y)$. On rappelle que $-1 \leq r(X,Y) \leq 1$, que la droite $D_{Y/X}$ "monte" si $r(X,Y) > 0$ et descend si $r(X,Y) < 0$, et que la corrélation entre Yet X est proche d'être linéaire si $r(X,Y)$ est proche de 1 ou de -1.

   Puis pour la droite $D_{X/Y}$:

   (c) CASIO : Dans la suite d'instructions précédentes, il suffit d'inverser les SETTINGS : SET→2Var XList : List 2, 2Var YList : List1, 2Var Freq : 1.
   TI : STAT→LinReg($ax+b$) L_2,L_1 (on inverse L_1 et L_2).

e) Une fois calculés les coefficients des différentes droites, on procède ainsi pour leur dessin.

   (a) Tout d'abord, on se souvient qu'il suffit de connaître deux points pour tracer une droite.

   (b) La *droite de Mayer* passe par les points G_1 et G_2.

   (c) La *droite de régression $D_{Y/X}$* passe par le barycentre G et par le point $(0, b)$.

(d) La *droite de régression* $D_{X/Y}$ passe le barycentre G et par le point $\left(\hat{b}, 0\right)$.

(e) Si l'abscisse 0 ou l'ordonnée 0 sont loin du cadre du dessin, et donc inutilisables, on utilise une abscisse x_0 ou une ordonnée y_0 raisonnables, dans le cadre du dessin, pour trouver un autre point l'aide de l'équation de la droite à tracer :
pour la droite $D_{Y/X}$ l'ordonnée correspondant à x_0 est $ax_0 + b_0$, et on a donc le point $(x_0, ax_0 + b_0)$.
pour la droite $D_{X/Y}$ l'abscisse correspondant à y_0 est $\hat{a}y_0 + \hat{b}_0$, et on a donc le point $\left(\hat{a}y_0 + \hat{b}_0, y_0\right)$.

La droite de régression de Mayer. Si la corrélation entre la variable X et la variable Y est de type linéaire, le problème est de trouver l'équation d'une droite autour de laquelle se regroupe le nuage de points donné par X et Y. Nous avons vu en cours comment trouver deux de ces droites possibles : les droites $D_{Y/X}$ et $D_{X/Y}$.

Il en existe une troisième : c'est la *droite de régression de Mayer*. Voici comment on la contruit. *Après avoir rangé les points du nuage par abscisses croissantes*, on sépare le nuage en deux groupes d'effectifs égaux : le groupe des premiers points et le groupe des dernier points (voir dans la correction de l'Exercice 4 ce qu'il faut faire si l'effectif du nuage de points est impair). On note $G_1 = (x_1, y_1)$ le barycentre du premier groupe de points, et $G_2 = (x_2, y_2)$ le barycentre du deuxième groupe de points. ***La droite de régression de Mayer est la droite passant par les points*** G_1 ***et*** G_2**.** Son équation est (voir formulaire) :

$$Y = a\,(X - x_1) + y_1, \text{ avec } a = \frac{y_2 - y_1}{x_2 - x_1}.$$

RÉPONSES AUX EXERCICES.

Exercice 1.

0. Le barycentre est $G = (m\,(X)\,, m\,(Y)) = (4.5, 23.375)$.

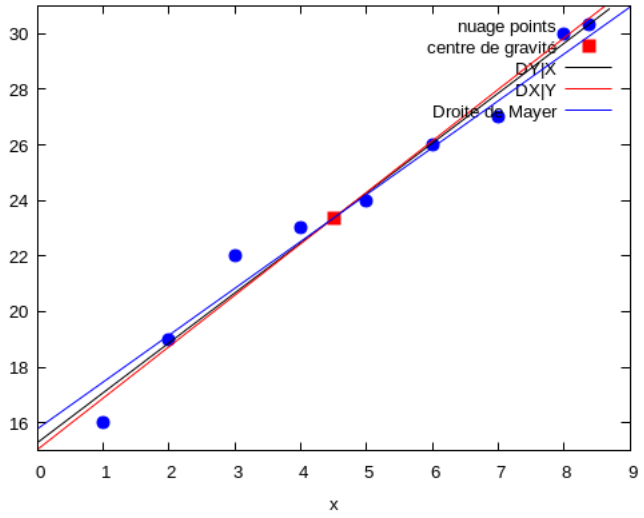

Figure 1: Exercice 1, Nuage de points et droites de régression

1. Pour déterminer l'équation de la *droite de Mayer*, on calcule déja le barycentre $G_1 = (x_1, y_1)$ de la première moitié de points (en l'occurrence les 4 premier points), puis le barycentre $G_2 = (x_2, y_2)$ de la seconde moitié de points (en l'occurrence les 4 derniers points). On trouve $G_1 = (2.5, 20.0)$ et $G_2 = (6.5, 26.75)$. On voit alors dans le formulaire l'équation $Y = a\,(X - x_1) + y_1$ avec $a = \dfrac{y_2 - y_1}{x_2 - x_1}$ de la *droite de régression de Mayer*, qui est la droite passant par les deux points G_1 et G_2. On trouve :

$$Y = 1.6875(X - 2.5) + 20.0.$$

Cette droite se dessine en joignant simplement les points G_1 et G_2.
Commentaire. Cette droite passe par le centre de gravité G du nuage de points.
Attention : le coefficient a qu'on calcule pour la droite de Mayer n'est pas le coefficient a qu'on calcule pour la droite de régression $D_{Y/X}$!
Si on utilise la fonction CALC du menu STAT de la calculatrice pour le calcul des points G_1 et G_2, il est recommandé de traiter la question sur la droite de Mayer en dernier. Pour calculer G_1, on peut recopier dans un premier temps les

listes L_1 et L_2 dans les listes L_3 et L_4, effacer la seconde moitié des points de ces deux listes, et calculer rapidement, à l'aide de la fonction STAT - 2-VAR *à laquelle on aura donné comme variables L_3 et L_4, les moyennes $\bar{x}$ et $\bar{y}$ de chacune des deux listes.*
On procède de même pour calculer G_2 en recopiant à nouveau L_1 et L_2 dans les listes L_3 et L_4, puis en supprimant cette fois la première moitié des points de ces deux listes.

2. $C(X,Y) = m(XY) - m(X)\,m(Y) = \frac{1}{8}\sum_{i=1}^{8} x_i y_i - m(X)\,m(Y) = \frac{917}{8} - 4.5 \times 23.375 = 9.437.$
3. Calcul des droites $D_{Y/X}$ et $D_{X/Y}$.
   (a) Droite $D_{Y/X}$. $a = \frac{C(X,Y)}{V(Y)} = \frac{9.437}{2.291^2} = 1.797$. $b = m(Y) - a \cdot m(X) = 23.375 - 1.797 \times 4.5 = 15.288$. L'équation de $D_{Y/X}$ est : $Y = aX + b = 1.797X + 15.286$.
   (b) Droite $D_{X/Y}$. $\hat{a} = \frac{C(X,Y)}{V(Y)} = \frac{9.547}{4.181^2} = 0.540$. $\hat{b} = m(X) - \hat{a} \cdot m(Y) = 4.5 - 0.540 \times 23.375 = -8.117$.
4. Chiffre d'affaire prévisible pour la 10ème année. On remplace simplement X par 10 dans les équations des droites.
   Avec la droite de Mayer, on a : $1.6875(10 - 2.5) + 20.0 = 32.656$.
   Avec la droite $D_{Y/X}$, on a : $1.7976 \times 10 + 15.285 = 33.26$.
5. Le coefficient de corrélation linéaire est : $r(X,Y) = \frac{C(X,Y)}{\sigma(X) \cdot \sigma(Y)} = \frac{9.437}{2.291 \times 4.181} = 0.985$. On voit qu'on a, conformément à la figure, un coefficient positif (et donc une droite $D_{Y/X}$ "montante"), très proche de 1 : on en conclut qu'on peut considérer que la corrélation entre X et Y est *forte*, de type *linéaire* (ce qui correspond bien à la figure)

Exercice 2.

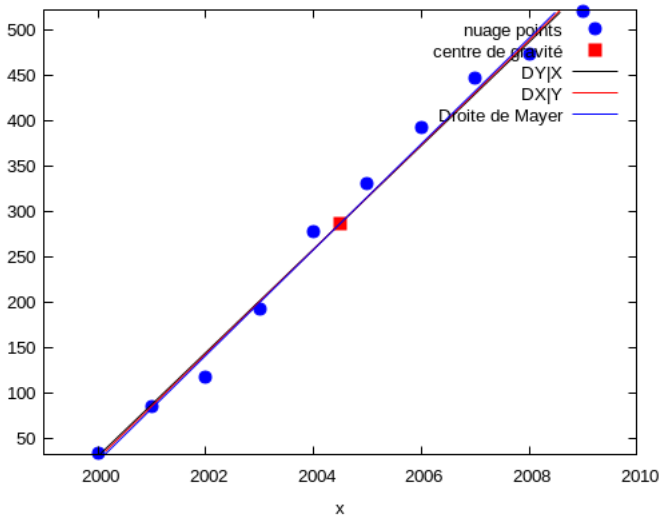

Figure 2: Exercice 2, Nuage de points et droites de régression

0. $G = (m(X), m(Y)) = (2004.5, 287.0)$.
1. $G_1 = (2002.0, 141.2)$ et $G_2 = (2007.0, 432.8)$, donc l'équation de la droite de Mayer est $Y = 58.32(X - 2002.0) + 141.2$.
2. $C(X,Y) = m(XY) - m(X)\,m(Y) = \frac{5757621}{10} - 2004.5 \times 287.0 = 470.6$.
3. $D_{Y/X} : Y = 57.042X - 1.1405 \times 10^5$; $D_{X/Y} : X = 0.0173Y + 1999.5$
4. Prévision pour $X = 2011$
   (a) avec Mayer : $58.32 \times (2011 - 2002.0) + 141.2 = 666.07$.
   (b) avec $D_{Y/X}$: $57.42 \times 2011 - 1.1405 \times 10^5 = 1421.6$
5. $r(X,Y) = \frac{C(X,Y)}{\sigma(X)\,\sigma(Y)} = 0.99492$. A nouveau, le coefficient est positif (droite "montante") et très proche de 1 : corrélation *forte* entre X et Y, de type *linéaire*.

Exercice 3.

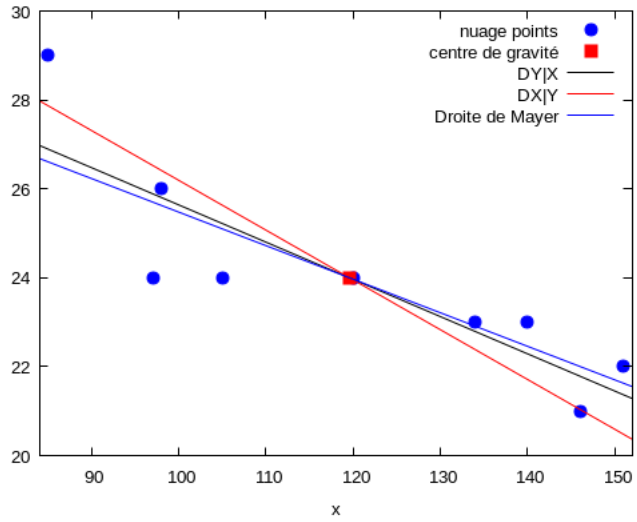

Figure 3: Exercice 3, Nuage de points et droites de régression

0) $G(X,Y) = (119.55, 24.0)$.
1) Pas très intéressant.
2) $C(X,Y) = m(XY) - m(X)\,m(Y) = \dfrac{25431}{9} - 119.55 \times 24.0 = -43.666$.
3) $D_{Y/X} : Y = -0.0837X + 34.009$, et $D_{X/Y} : -8.9318Y + 333.91$.
4) Avec $X = 125$ et $D_{Y/X}$, on a : $-0.083719 \times 125 + 34.009 = 23.544$.
5) $r(X,Y) = \dfrac{C(X,Y)}{\sigma(X)\,\sigma(Y)} = -0.86473$. Négatif, donc droite "descendante", et moins proche de -1 que les exercices précédents : nous dirons donc corrélation *relativement forte* entre X et Y, de type linéaire. On voit d'ailleurs bien sur la figure que, alors que les droites de régressions étaient presque confondues dans les exercices précédents, elles sont plus espacées dans cet exercice.

Exercice 4. Deux remarques sur cet exercice.

a) Typiquement, *il ne faut pas oublier de ranger* la liste des pourcentages de population agricoles (qui ne sont pas données par ordre croissant dans l'énoncé), et donc ordonner de façon cohérente la liste des calories/jour.

b) Ici, il y a un *nombre impair de points* dans le nuage de points. Comment donc séparer les points en deux moitiés ? L'usage est de regarder, sur le nuage de points, comment le point central est positionné par rapport aux deux moitiés restantes. Il est en général plus proche d'une des moitiés que de l'autre. On considère donc qu'il "fait partie" de cette moitié là, et qu'on a ainsi "une moitié plus grosse que l'autre".

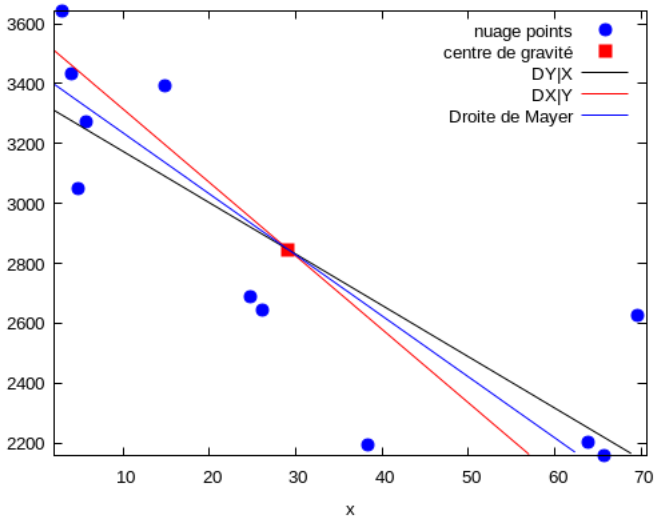

Figure 4: Exercice 4, Nuage de points et droites de régression

0. $G(X,Y) = (m(X), m(Y)) = (29.154, 2845.7)$.
1. Pas très intéressant.
2. L'effectif est de $n = 11$ points, c'est à dire qu'il y a *un nombre impair de points.* Or l'examen du nuage de points, ou de la liste des points après rangement, montre que le point central de la liste de points, qui est le 6ème après rangement, de coordonnées $(24.7, 2687)$, est beaucoup plus proche du point suivant $(26.2, 2643)$ que du point précédent $(14.8, 3394)$. C'est très net sur la figure, sur laquelle on voit bien les deux points très proches au milieu. On sépare donc en deux groupes de points : le premier groupe contient uniquement les 5 premiers points, et le second groupe contient également le point central, contient donc les 6 derniers points.
   Le barycentre des 5 premiers points est $G_1 = (9.516, 3246.1)$, et le barycentre des 6 derniers points est $G_2 = (52.72, 2365.2)$. La droite de Mayer a donc pour équation :
$$Y = -20.391 \times (X - 9.516) + 3246.1.$$
3. $C(X,Y) = m(XY) - m(X)\,m(Y) = \dfrac{7.9342 \times 10^5}{11} - 29.154 \times 2845.7 = -1.0836 \times 10^4 = -10836.$
4. $D_{Y/X} : Y = -17.163X + 3346.1$
5. $r(X,Y) = \dfrac{C(X,Y)}{\sigma(X)\,\sigma(Y)} = -\dfrac{1.0836 \times 10^4}{25.127 \times 515.81} = -0.8361$. Ce cofficient est négatif, pas trop éloigné de -1, on a donc une corrélation relativement linéaire entre X et Y qui correspond à une droite décroissante. Ce résultat intéressant indique donc que, de façon statistique, *plus la proportion de population agricole dans un pays est grande, moins ses habitants consomment de calories par jour !*