M1 MEEF 2016-2017
Séquence 4

Correction de l'exercice 3 de la feuille 4 sur la statistique.

Enoncé

Dans une ferme industrielle, le service vétérinaire veut modifier le régime alimentaire des vaches, dans le but d'augmenter la production laitière. Pour cela, on a choisi au hasard 15 vaches que l'on a nourries pendant un mois avec l'aliment habituel et l'on a relevé pour chaque vache X la production quotidienne moyenne de lait exprimée en kg. Puis, on a nourri ces mêmes vaches pendant un mois avec le nouvel aliment et on a relevé de même Y la production quotidienne moyenne de chaque vache.

N° de la vache	1	2	3	4	5	6	7	8	9	10	11
X en kg/jour	27,6	23,4	25,2	28,2	28,8	25,8	27	27	29,4	28,2	30
Y en kg/jour	28,8	25,6	26,4	28	31,2	27,2	28,8	28	29,6	29,2	28,4

N° de la vache	12	13	14	15
X en kg/jour	28,2	32,4	29,4	30
Y en kg/jour	29,6	31,2	32	29,2

On note x_i la production de la vache i avec l'aliment habituel et y_i avec l'aliment nouveau. On note également $N = 15$, X une variable aléatoire de loi $\frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}$ et Y de loi $\frac{1}{N}\sum_{i=1}^{N}\delta_{y_i}$.

1. Calculer $\mathbb{P}(X = x_1)$ et $\mathbb{P}(Y = y_1)$. Quel est le terme statistique pour désigner $\mathbb{P}(X = 27,6)$?
2. On pose $U = \frac{X-28,2}{0,6}$, $u_i = \frac{x_i-28,2}{0,6}$, $V = \frac{Y-28,8}{0,8}$ et $v_i = \frac{y_i-28,8}{0,8}$. On sait que

$$\sum_{i=1}^{15} u_i = -4; \qquad \sum_{i=1}^{15} u_i^2 = 190; \qquad \sum_{i=1}^{15} v_i = 1,5; \qquad \sum_{i=1}^{15} v_i^2 = 67,75; \qquad \sum_{i=1}^{15} u_i v_i = 91.$$

   Exprimer ces sommes en termes probabilistes de U et de V.
3. Écrire X et Y en fonction de U et V et en déduire $\mathbb{E}(X)$, $\mathbb{E}(Y)$, $\mathrm{Var}(X)$, $\mathrm{Var}(Y)$, $\mathrm{Cov}(X, Y)$.
4. Calculer la valeur moyenne et l'écart type des $(x_i)_{1\leqslant i\leqslant 15}$ et des $(y_i)_{1\leqslant i\leqslant 15}$.
5. Calculer le coefficient de corrélation r entre $(x_i)_{1\leqslant i\leqslant 15}$ et $(y_i)_{1\leqslant i\leqslant 15}$, que peut-on en conclure ?

Solution

1. Puisque $x_1 = 27,6$ et que seule la vache 1 produit cette quantité de $27,6$ kg/jour, on a directement $\mathbb{P}(X = x_1) = \frac{1}{N} = \frac{1}{15}$. Pour $y_1 = 28,8$, on note que la vache 7 produit aussi la même quantité de lait, $y_7 = y_1$. Donc $\mathbb{P}(X = y_1) = \frac{2}{N} = \frac{2}{15}$. Le terme $\mathbb{P}(X = 27,6)$ est la fréquence de la valeur $27,6$.

2. Les u_i, pour $i \in \{1, \dots, 15\}$, représentent toutes les valeurs possibles de la variable U avec redondance : les u_i ne sont pas nécessairement distincts. On a donc

$$\sum_{i=1}^{15} u_i = N \times \frac{1}{N} \sum_{i=1}^{15} u_i = N\mathbb{E}(U).$$

   De même,

$$\sum_{i=1}^{15} u_i^2 = N\mathbb{E}\left(U^2\right), \qquad \sum_{i=1}^{15} v_i = N\mathbb{E}(V), \qquad \sum_{i=1}^{15} v_i^2 = N\mathbb{E}\left(V^2\right).$$

   Attention le dernier terme est sensiblement plus délicat. On rappelle que

$$\mathbb{E}(UV) = \sum_{1 \leqslant i,j \leqslant N} u_i v_j \mathbb{P}(U = u_i, V = v_j).$$

   Maintenant, il faut supposer exactement comme dans l'exercice 1 que les variables aléatoire X et Y sont les marginales d'une variable aléatoire $M = (X, Y)$ dont les seuls valeurs possibles sont (x_i, y_i) :

$$M \sim \frac{1}{N} \sum_{i=1}^{N} \delta_{(x_i, y_i)}.$$

   Notamment $\mathbb{P}(X = x_i, Y = y_j) = 0$ si $i \neq j$ et $\mathbb{P}(X = x_i, Y = y_i) = 1/N$ pour tout $i \in \{1, \dots, N\}$. Et donc

$$\begin{aligned}\mathbb{P}(U = u_i, V = v_j) &= \mathbb{P}(X = 0,6u_i + 28,2, Y = 0,8v_j + 28,8) \\ &= \mathbb{P}(X = x_i, Y = y_j) = \frac{\delta_{i,j}}{N}\end{aligned}$$

   où $\delta_{i,j}$ est le symbole de Kronecker et vaut 0 si $i \neq j$ et 1 sinon. On en déduit que

$$\mathbb{E}(UV) = \frac{1}{N} \sum_{i=1}^{N} u_i v_i$$

3. X et Y sont des transformations affines de U et V :

$$X = 0,6U + 28,2 \qquad \text{et} \qquad Y = 0,8V + 28,8.$$

   Donc

$$\begin{aligned}\mathbb{E}(X) &= 0,6\mathbb{E}(U) + 28,2 = \frac{0,6}{15}(-4) + 28,2 = 28,4 \\ \mathbb{E}(Y) &= 0,8\mathbb{E}(V) + 28,8 = \frac{0,8}{15}(1,5) + 28,8 = 28,88\end{aligned}$$

et

$$\mathrm{Var}(X) = (0,6)^2\,\mathrm{Var}(U) = 0,36\left(\mathbb{E}(U^2) - \mathbb{E}(U)^2\right) = 0,36\left(\frac{190}{15} - \left(\frac{-4}{15}\right)^2\right) = 4,53$$

$$\mathrm{Var}(Y) = (0,8)^2\,\mathrm{Var}(V) = 0,64\left(\mathbb{E}(V^2) - \mathbb{E}(V)^2\right) = 0,64\left(\frac{67,75}{15} - \left(\frac{1,5}{15}\right)^2\right) = 2,88.$$

Enfin,

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathrm{Cov}\,(0,6U + 28,2\,;\,0,8V + 28,8)\\ &= 0,6 \times 0,8\,\mathrm{Cov}\,(U\,;\,V)\\ &= 0,48\,(\mathbb{E}\,(UV) - \mathbb{E}(U)\mathbb{E}(V))\\ &= 0,48\left(\frac{91}{15} - \frac{-4}{15} \times \frac{1,5}{15}\right)\\ &= 2,92.\end{aligned}$$

4. Les moyennes c'est déjà fait : $\mathbb{E}(X) = 28,4$ et $\mathbb{E}(Y) = 28,88$ et pour les écarts-types il suffit de prendre les racines carrées des variances : $\sqrt{\mathrm{Var}(X)} = 2,13$ et $\sqrt{\mathrm{Var}(Y)} = 1,7$.

5. On en déduit le coefficient de corrélation

$$r = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \frac{2,92}{2,13 \times 1,7} = 0,81.$$

Conclusion : pas si mal, r n'est pas trop trop loin de 1 la régression linéaire semble approcher raisonnablement le nuage de points.