Université de Carthage
Ecole Supérieure de la Statistique
et de l'Analyse de l'Information à Tunis

Année Universitaire
2017-2018
Première Année

Statistique Descriptive
Session Principale: janvier 2018

(Aucun document autorisé)
(2 pages)

Enseignante : Mme Héla Ouaili-Mallek *Durée : 1h30*

Exercice 1 *Afin de recruter deux étudiants saisonniers, une entreprise décide de soumettre chacun des* 10 *candidats qui se sont présentés à une épreuve écrite de culture générale et un entretien (notés sur 20). Les notes sont consignées dans le tableau suivant :*

Epreuve	*8*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
Entretien	*6*	*7*	*10*	*6*	*11*	*9*	*10*	*12*	*11*	*13*

Nous avons préalablement procédé à quelques calculs et obtenu les résultats qui suivent:

$\sum_i x_i = 116$, $\sum_i y_i = 95$, $\sum_i x_i^2 = 1420$, $\sum_i y_i^2 = 957$ *et* $\sum_i x_i y_i = 1154$

1. *On souhaite vérifier l'existence d'une relation linéaire entre les deux notes. Quelle variable endogène choisir? (justifier)*

2. *Peut-on envisager l'existence d'une telle relation? (justifier)*

3. *On suppose que cette relation linéaire est effective. Evaluer son intensité et donner sons sens.*

4. *Donner l'équation de la droite d'ajustement des notes de l'entretien oral sur les notes de l'épreuve écrite.*

5. *En déduire l'équation de la droite d'ajustement des notes de l'épreuve écrite sur les notes de l'entretien oral.*

6. *Calculer le coefficient de détermination du modèle ajustant les notes de l'entretien oral sur les notes de l'épreuve écrite. Commenter.*

Exercice 2 *Les données qui suivent résument la répartition des salariés d'une jeune start-up selon l'âge et l'ancienneté dans l'entreprise en nombre d'années:*

Age / *Ancienneté*	[24 ; 26[	[26 ; 28[	[28 ; 30[
0	2	1	0
1	1	2	1
2	0	2	1
3	2	1	0
4	1	2	1
5	0	2	1

1. *Peut-on confirmer l'existence d'une liaison fonctionnelle entre les deux variables? (justifier)*

2. *Qu'en est-il de l'indépendance de ces deux variables? (justifier)*

3. *Calculer la distance du khi2. Interpréter le résultat. Pouvait-on s'y attendre? (justifier)*

4. *On s'intéresse à la* ***distribution conditionnelle de l'ancienneté des salariés sachant qu'ils sont âgés entre 24 et 26 ans****.*

   (a) *Donner la distribution statistique.*

   (b) *Déterminer et représenter graphiquement la fonction de répartition.*

   (c) *En déduire la médiane.*

   (d) *Proposer le plus petit intervalle contenant la médiane et qui représente au moins 30% des effectifs.*

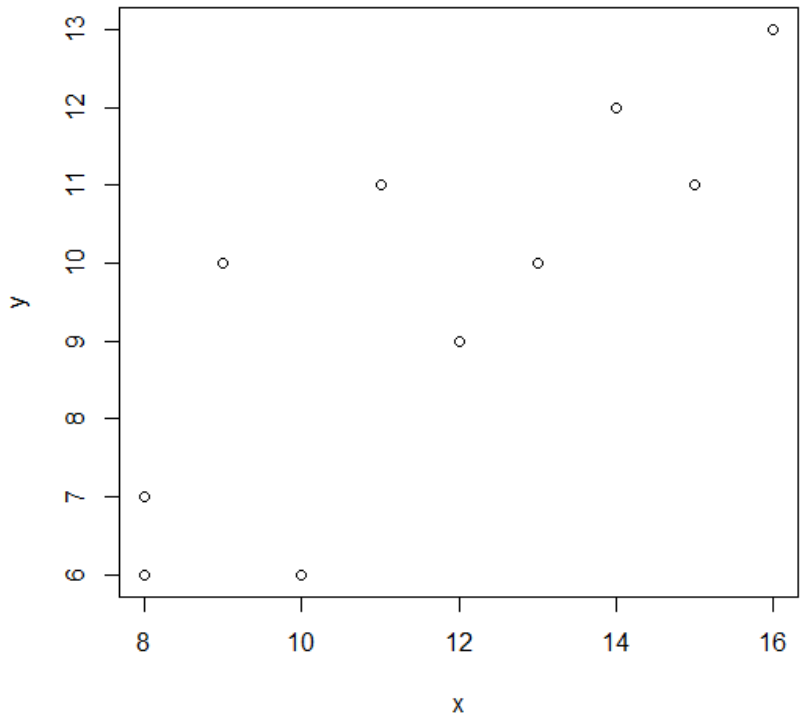
Corrigé exercice 1

Epreuve	8	8	9	10	11	12	13	14	15	16
Entretien	6	7	10	6	11	9	10	12	11	13

$\sum_i x_i = 116$ $\sum_i y_i = 95$ $\sum_i x_i^2 = 1420$ $\sum_i y_i^2 = 957$ $\sum_i x_i y_i = 1154$

1. Le choix de la variable endogène repose sur l'existence de risque d'erreur de mesure ou d'observation. Or les notes des épreuves écrites sont généralement moins subjectives que celles des épreuves orales. Le risque d'erreur de mesure est donc plus élevé pour ces dernières et nous ajusterons les notes de l'entretien (y) sur es notes de l'épreuve écrite (x).

2. Pour envisager une relation linéaire, il faut que le nuage de points dessine une tendance linéaire. Ce qui semble être le cas.



Nuage des points des notes obtenues

3. L'intensité de la relation linéaire se mesure à travers le coefficient de corrélation linéaire:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \qquad s_{xy} = \frac{1}{n}\sum_i x_i y_i - \overline{x}\ \overline{y}$$

$$\overline{x} = 11.6 \qquad \overline{y} = 9.5 \qquad s_{xy} = 115.4 - 11.6 * 9.5 = 5.2$$

$$s_x^2 = \frac{1}{n}\sum_i x_i^2 - \overline{x}^2 = 142 - 11.6^2 = 7.44 = (2.73)^2$$

$s_y^2 = \frac{1}{n}\sum_i y_i^2 - \overline{y}^2 = 95.7 - 9.5^2 = 5.45 = (2.33)^2$

$r_{xy} = \frac{5.2}{2.73 * 2.33} = 0.82$

Nous pouvons conclure que les deux notes sont corrélées positivement.

4. $\widehat{y} = \widehat{a}\ x + \widehat{b}$

   $\widehat{a} = \frac{s_{xy}}{s_x^2} = \frac{5.2}{7.44} = 0.7 \qquad \widehat{b} = \overline{y} - \widehat{a}\ \overline{x} = 9.5 - 0.7 * 11.6 = 1.38$

   $$\widehat{y} = 0.7\ x + 1.38$$

5. $\widehat{x} = \widehat{a}'\ y + \widehat{b}'$

   $r_{xy}^2 = \widehat{a} * \widehat{a}' \Longrightarrow \widehat{a}' = \frac{(0.82)^2}{0.7} = 0.96 \Longrightarrow \widehat{b}' = 11.6 - 0.96 * 9.5 = 2.48$

   $$\widehat{x} = 0.96\ y + 2.48$$

6. $R^2 = \frac{s_{\widehat{y}}^2}{s_y^2} = \frac{\widehat{a}^2 s_x^2}{s_y^2} = \frac{0.7^2 * 7.44}{5.45} = 0.67$

7. Le modèle n'explique que 67% de la variance totale. Autrement dit 33% de l'information est perdue!

Corrigé de l'exercice 2:

1. Les 2 variables ne sont pas liées fonctionnellement puisqu'on a $n_{11} = 2$ et $n_{12} = 1$, tous deux différents de zéro et $n_{11} = 2$ et $n_{21} = 1$, différents de zéro.

2. Les 2 variables ne sont pas non plus indépendantes puisque $f_{11} = 0.1 \neq f_{1\cdot} * f_{\cdot 1} = \frac{6}{20} * \frac{3}{20}$

3. $\sum\sum \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} = 1.35$

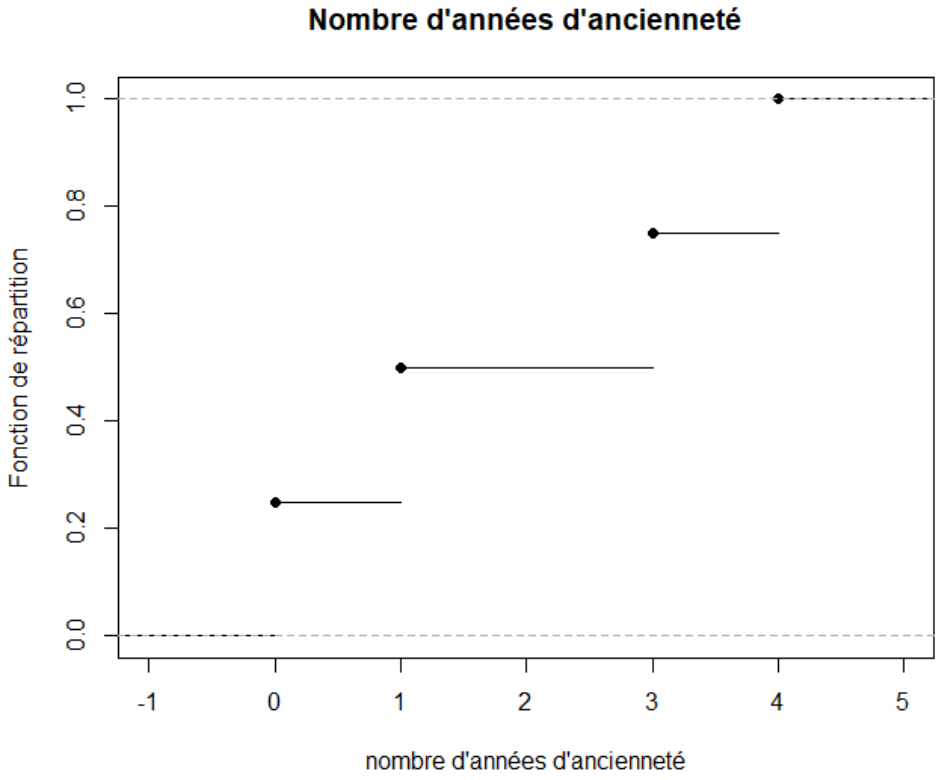
   $D^2 = n\left(\sum\sum \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} - 1\right) = 20\,(1.35 - 1) = 7.$

   On a $\min\left(n\,(p-1), n\,(k-1)\right) = \min\left(20 * 5, 20 * 2\right) = 40 >> 7 > 0.$

   Nous pouvons conclure sans grand risque d'erreur que les deux variables ne sont pas indépendantes.

   (a) Fonction de répartition

Ancienneté	0	1	3	4
f_i^1	0.33	0.17	0.33	0.17
$F\,(x_i)$	0	0.33	0.5	0.83



(b) Médiane: $F(3) = 0.5 \Longrightarrow M_e = 3$

Ancienneté	0	1	3	4
f_i^1	0.33	0.17	0.33	0.17
$F(x_i)$	0	0.33	0.5	0.83

(c) $F(1) = 0.33, F(3) = 0.5 \Longrightarrow x_{0.35} = 1$

$F(3) = 0.5, F(4) = 0.83 \Longrightarrow x_{0.65} = 3$

L'intervalle $[1\ ;\ 3]$ contient la médiane et au moins 30% des effectifs.