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

Année Universitaire 2013-2014
Première Année

Statistique Descriptive
Examen final
janvier 2014

Enseignante : Mme Héla Ouaili Mallek

Durée :1h30

(02 pages)

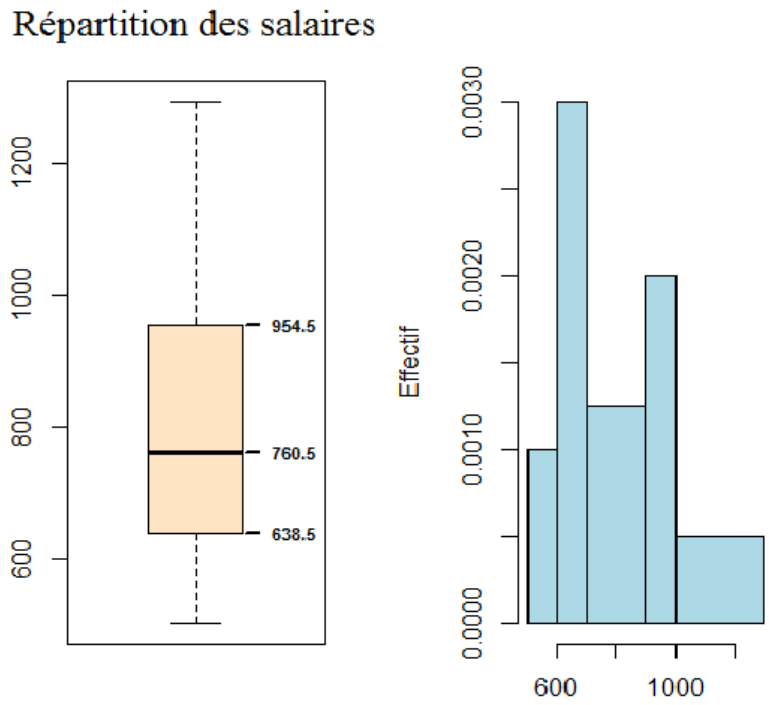
Exercice 1 *On s'intéresse à deux variables statistiques quantitatives X et Y pour lesquelles certains calculs préalables ont été effectués:*

$$\sum_{i=1}^{10} x_i = 56 \qquad \sum_{i=1}^{10} x_i^2 = 384 \qquad \sum_{i=1}^{10} y_i = 44 \qquad \sum_{i=1}^{10} y_i^2 = 238$$

$$\sum_{i=1}^{10} x_i\ y_i = 195 \qquad \sum_{i=1}^{10} u_i^2 = 0.69.$$

1. *Calculer les variances empiriques de ces deux variables.*
2. *Calculer la covariance entre X et Y.*
3. *L'objectif est de mettre en évidence une relation de linéarité entre X et Y. Donner l'équation de la droite de régression de Y sur X.*
4. *Calculer le coefficient de corrélation empirique entre X et Y. Commenter.*
5. *A partir des résultats précédents, déduire l'expression de la droite de régression de X sur Y.*
6. *Laquelle de ces deux droites choisiriez-vous? Justifier.*

Exercice 2 *Nous avons représenté graphiquement la distribution des 200 employés d'une entreprise selon le salaire mensuel exprimé en dinars.*



1. *Donner, à partir de ces graphiques, une description de cette distribution.*

2. *Après avoir regroupé les salaires en classes, nous avons consigné la répartition de ces salaires dans le tableau ci-après.*

Salaire	*Effectif*
$[500, 600[$	20
$[600, 700[$	60
$[700, 900[$	50
$[900, 1000[$	40
$[1000, 1300[$	30

   Calculer le salaire modal, le salaire médian et le salaire moyen.

3. *Que peut-on conclure quant à la forme de la distribution?*

4. *Déterminer, par interpolation linéaire, le premier et le troisième quartile. Ces valeurs étaient-elles prévisibles?*

5. *On s'intéresse maintenant à l'étude de l'inégalité de la répartition des salaires. Construire la courbe de Lorenz.*

6. *Calculer la médiale. Commenter.*

7. *Calculer l'indice de Gini. Interpréter le résultat.*

Corrigé de l'exercice 1 *:*

$$\sum_{i=1}^{10} x_i = 56 \qquad \sum_{i=1}^{10} x_i^2 = 384 \qquad \sum_{i=1}^{10} y_i = 44 \qquad \sum_{i=1}^{10} y_i^2 = 238$$
$$\sum_{i=1}^{10} x_i\ y_i = 195 \qquad \sum_{i=1}^{10} u_i^2 = 0.69.$$

1. $\overline{x} = \dfrac{1}{10}\sum_{i=1}^{10} x_i = 5.6 \qquad s_X^2 = \dfrac{1}{10}\sum_{i=1}^{10} x_i^2 - \overline{x}^2 = 38.4 - 5.6^2 = 7.04$

   $\overline{y} = \dfrac{1}{10}\sum_{i=1}^{10} y_i = 4.4 \qquad s_Y^2 = \dfrac{1}{10}\sum_{i=1}^{10} y_i^2 - \overline{y}^2 = 23.8 - 4.4^2 = 4.44$

2. $s_{XY} = \dfrac{1}{10}\sum_{i=1}^{10} x_i\ y_i - \overline{x}\ \overline{y} = 19.5 - 5.6 * 4.4 = -5.14$

3. $\widehat{y_i} = \widehat{a}\ x_i + \widehat{b} \qquad \widehat{a} = \dfrac{s_{XY}}{s_X^2} = -\dfrac{5.14}{7.04} = -0.730\,11$

   $\widehat{b} = \overline{y} - \widehat{a}\ \overline{x} = 4.4 + 0.73011 * 5.6 = 8.4886$

   $\widehat{y_i} = -0.730\,11 * x_i + 8.4886$

4. $r_{XY} = \dfrac{s_{XY}}{s_X\ s_Y} = -\dfrac{5.14}{\sqrt{7.04}\sqrt{4.44}} = -0.91936$ *On a donc une forte corrélation négative entre les deux variables X et Y.*

5. $\widehat{x_i} = \widehat{a'}\ y_i + \widehat{b'} \qquad \widehat{a'} * \widehat{a} = r_{XY}^2 \Longrightarrow \widehat{a'} = \dfrac{r_{XY}^2}{\widehat{a}} = \dfrac{(0.91936)^2}{-0.73011} = -1.1577$

   $\widehat{b'} = \overline{x} - \widehat{a'}\overline{y} = 5.6 + 1.1577 * 4.4 = 10.694$

   $\widehat{x_i} = -1.1577 * y_i + 10.694$

6. $\dfrac{s_{\widehat{Y}}^2}{s_Y^2} == \dfrac{\widehat{a}^2 s_X^2}{4.44} = \dfrac{(0.73011)^2 * 7.04}{4.44} = 0.84521$

   $\dfrac{s_{\widehat{X}}^2}{s_X^2} = \dfrac{\widehat{a'^2} * s_Y^2}{s_X^2} = \dfrac{(1.1577)^2 * 4.44}{7.04} = 0.84528$

   $\dfrac{s_{\widehat{X}}^2}{s_X^2} > \dfrac{s_{\widehat{Y}}^2}{s_Y^2}$ *Donc la régression de X sur Y est préférable à celle de Y sur X.*

Corrigé de l'exercice 2 *:*

1. *L'histogramme exprime une hétérogénéïté à travers une bimodalité. Cette hétérogénéïété pourrait refléter, par exemple, deux catégories de salariés, les ouvriers et employés, d'une part et les cadres, d'autre part. Par ailleurs, on peut discerner une dissymétrie de la distribution salariale avec une plus forte concentration des bas salaires.*

Cette dissymétrie se cofirme par le box-plot avec une médiane appartenant à la moitié inférieure de la boîte et une moustache inférieure de longueur inférieure de moitié à la longueur de la moustache supérieure.

2. *Calculs intermédiaires:*

$Salaire$	n_i	n_i^c	$f(e_i)$	$F(e_i)$	$f_i\ c_i$	$\sum_{j=1}^{i} f_j\ c_j\ /\overline{x}$
$[500, 600[$	20	20	0.1	0.1	55	0.06769
$[600, 700[$	60	60	0.3	0.4	195	0.30769
$[700, 900[$	50	25	0.25	0.65	200	0.55385
$[900, 1000[$	40	40	0.2	0.85	190	0.78769
$[1000, 1300[$	30	10	0.15	1	172.5	1
$Total$	200	–	1	–	812.5	

Salaire modal : *Amplitude de référence* $a = 100$. *Classe modale :* $[600, 700[$

$$M_o = 600 + 100 * \frac{60 - 20}{(60 - 20) + (60 - 25)} = 653.33 dinars.$$

Salaire médian : *Classe médiane :* $[700, 900[$

$$M_e = 700 + 100 * \frac{0.5 - 0.4}{0.65 - 0.4} = 740 dinars$$

Salaire moyen :

$$\overline{x} = \sum_{i=1}^{5} f_i\ c_i = 0.1 * 550 + 0.3 * 650 + 0.25 * 800 + 0.2 * 950 + 0.15 * 1150 = 812.5 dinars.$$

3. *Malgré les inégalités* $M_o \leq M_e \leq \overline{x}$, *il n'est pas pertinent de conclure quant à la forme de la distribution et ce, en raison de la bi-modalité de la distribution.*

4. $Q_1 \in [600, 700[$

$$Q_1 = 600 + 100 * \frac{0.25 - 0.1}{0.4 - 0.1} = 650 dinars$$

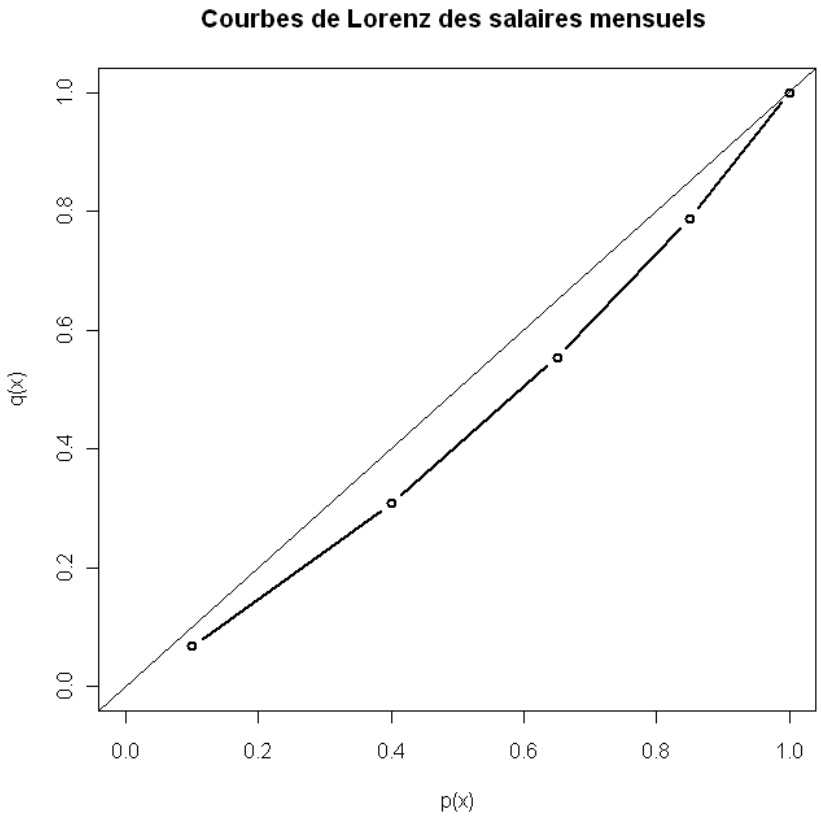
$Q_2 \in [900, 1000[$

$$Q_2 = 900 + 100 * \frac{0.75 - 0.65}{0.85 - 0.65} = 950 dinars$$

*Ces valeurs ne s'éloignent pas des vraies valeurs figurant sur le box-plot (*638.5 *pour le premier quartile et* 954.5 *pour le troisième quartile) ; elles correspondent donc à nos attentes.*

5. *On a* $p(e_i) = F(e_i)$ *et* $q(e_i) = \dfrac{\sum_{j=1}^{i} f_j\ c_j}{\overline{x}}$.

La courbe de Lorenz joint les points $(p(e_i), q(e_i))$.



6. *La médiale M_l appartient à* $[700, 900[$

   $$M_l = 700 + 100 * \frac{0.5 - 0.30769}{0.55385 - 0.30769} = 778.12 dinars$$

   La médiale est supérieure à la médiane ($740 dinars$*) mais la différence n'est pas très importante, ce qui laisse supposer que l'inégalité n'est pas très importante.*

7. *Indice de Gini :*

8. $$G = 1 - \sum_{i=1}^{5} \left(p\left(e_i\right) - p\left(e_{i-1}\right)\right)\left(q\left(e_i\right) + q\left(e_{i-1}\right)\right)$$

   $$G = 1 - \left(\begin{array}{c} 0.1 * 0.06769 + 0.3 * (0.30769 + 0.06769) + 0.25 * (0.55385 + 0.30769) \\ +0.2 * (0.55385 + 0.78769) + 0.15 * (1 + 0.78769) \end{array} \right)$$

   $$G = 0.128\,77$$

9. *l'indice de Gini n'est pas très élevé, ce qui confirme l'allure de la courbe de Lorenz et la valeur de la médiale : la distribution est peu inégalitaire.*