Corrigé du DM 1 - Indicateurs de dispersion paramétriques

Tran Viet Chi

20 novembre 2006

8.1 L'écart-type $\sigma(x)$ est la moyenne quadratique des distances des observations à la moyenne :

$$\sigma(x) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n}\sum_{j=1}^{k} n_j\,(x_j - \bar{x})^2}.$$

où l'on dénote par n le nombre d'individus et par k le nombre de classes et par n_j le nombre d'observations prenant la modalité x_j (effectif de la classe j). On rappelle pour mémoire les formules :

$$\sigma(x) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i^2) - \bar{x}^2} = \sqrt{\frac{1}{n}\sum_{j=1}^{k} n_j\,(x_j)^2 - \bar{x})^2}.$$

L'écart absolu moyen par rapport à la moyenne (arithmétique) $e(x)$ est la moyenne (arithmétique) des distances des observations à la moyenne :

$$e(x) = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|.$$

L'écart absolu moyen par rapport à la médiane $\widetilde{e}(x)$ est la moyenne (arithmétique) des distances des observations à la médiane :

$$\widetilde{e}(x) = \frac{1}{n}\sum_{i=1}^{n}|x_i - x_{med}|,$$

où x_{med} est la médiane.

Par l'inégalité de Cauchy-Schwarz (la moyenne arithmétique est plus petite que la moyenne quadratique) on a :

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

ce qui implique que $e(x) \leq \sigma(x)$.

Montrons que $\widetilde{e}(x) \leq e(x)$. Supposons pour simplifier que le nombre d'observations est pair.

$$\begin{aligned}
\widetilde{e}(x) =& \frac{1}{n}\sum_{i=1}^{n}|x_i - x_{med}| = \frac{1}{n}\sum_{i/\,x_i \geq x_{med}}|x_i - x_{med}| + \frac{1}{n}\sum_{i/\,x_i < x_{med}}|x_i - x_{med}| \\
=& \frac{1}{n}\sum_{i/\,x_i \geq x_{med}}(x_i - x_{med}) + \frac{1}{n}\sum_{i/\,x_i < x_{med}}(x_{med} - x_i) \\
=& \frac{1}{n}\sum_{i/\,x_i \geq x_{med}}((x_i - \bar{x}) + (\bar{x} - x_{med})) + \frac{1}{n}\sum_{i/\,x_i < x_{med}}((x_{med} - \bar{x}) + (\bar{x} - x_i)) \\
=& \frac{1}{n}\sum_{i/\,x_i \geq x_{med}}(x_i - \bar{x}) + \frac{1}{n}\sum_{i/\,x_i < x_{med}}(\bar{x} - x_i) + \underbrace{\frac{1}{n}\sum_{i/\,x_i \geq x_{med}}(\bar{x} - x_{med})}_{n/2 \text{ termes}} + \underbrace{\frac{1}{n}\sum_{i/\,x_i < x_{med}}(x_{med} - \bar{x})}_{n/2 \text{ termes}} \\
=& \frac{1}{n}\sum_{i/\,x_i \geq x_{med}}(x_i - \bar{x}) + \frac{1}{n}\sum_{i/\,x_i < x_{med}}(\bar{x} - x_i) \leq \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| = e(x).
\end{aligned}$$

8.2 Le coefficient de variation : $c(x) = \sigma(x)/\bar{x} = 0,48$. Pour l'indice de Kuznets, en utilisant les données de la question 4 :

$$K = \frac{1}{2n\bar{x}} \sum_{i=1}^{n} |x_i - \bar{x}| = \frac{1}{2n\bar{x}} 1000 \sum_{i=1}^{n} |y_i - \bar{y}| = \frac{1000 \times 766,68}{2 \times 120 \times 18935} = 0,17.$$

8.3.a) Le coefficient de variation pour les hommes est 0,48, et pour les femmes 0,38. Le coefficient de variation plus étendu pour les hommes que pour les femmes indique qu'il y a une plus grande dispersion relative des salaires chez hommes que chez les femmes.
8.3.b) Pour la moyenne :

$$\bar{x} = \frac{n_1}{n_1 + n_2}\bar{x}_1 + \frac{n_2}{n_1 + n_2}\bar{x}_2 = \frac{81}{120} \times 20596,30 + \frac{39}{120} \times 15484,62 = 18935.$$

Pour l'écart-type, on utilise la formule de la décomposition de la variance (question 5.1) qui nous indique comment retrouver la variance globale à partir des variances de chaque classe :

$$\begin{aligned}
\sigma(x) =& \sqrt{V(x)} = \sqrt{V_{inter} + V_{intra}} \\
=& \sqrt{\frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2}{n_1 + n_2} + \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2}} \\
=& \sqrt{\frac{81(20596,30 - 18935)^2 + 39(15484,62 - 18935)^2}{120} + \frac{81 \times 9928,75^2 + 39 \times 5941,14^2}{120}} = 9151,24
\end{aligned}$$