Probabilité et statistique I (partim statistique descriptive)
Bachelier en sciences informatiques
Vendredi 10 juin 2016

Théorie

Voir syllabus du cours.

Exercices

1. (a) Dénotons par $J = 4$ le nombre de classes. En utilisant que

$$1 = \sum_{i=1}^{J} f_i = 0.17 + f_2 + 0.28 + 0.02$$

et en résolvant cette équation d'inconnue f_2, on obtient $f_2 = 0.53$. Pour e_3, on utilise la définition de la moyenne (pour des observations groupées)

$$-3.9 = \bar{x} = \sum_{i=1}^{J} \frac{e_i + e_{i-1}}{2} f_i = \frac{-50-25}{2}0.17 + \frac{-25+0}{2}0.53 + \frac{0+e_3}{2}0.28 + \frac{160+e_3}{2}0.02$$

et en résolvant cette équation d'inconnue e_3, on trouve $e_3 = 50$.

(b) Les fréquences cumulées des 4 classes sont données successivement par 0.17, 0.7, 0.98 et 1.

(c) Voir Figure 1 (où la nouvelle ogive correspond à la ligne brisée construite à partir des symboles $\circ$).

2. (a) Indiçons respectivement Baltimore et Boston par 1 et 2. Trois paramètres de dispersion différents sont calculables à partir de l'énoncé:
   - L'écart-type s (ou la variance s^2): $s_1 = 33.92$ et $s_2 = 65.99$
   - L'écart inter-quartile $Q_3 - Q_1$: $EIQ_1 = 25.25$, $EIQ_2 = 31$
   - L'étendue $Max - Min$: $E_1 = 207$, $E_2 = 525$

Toutes ces informations nous amènent à la conclusion que les données sont clairement plus dispersées dans la série de Boston que dans celle Baltimore.

(b) Comme nous avons deux sous-populations, on peut utiliser le deuxième terme de la décomposition de la variance qui explique la variabilité entre les groupes (n étant l'effectif total):

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{2} n_i(\bar{x}_i - \bar{x})^2 \\
&= \frac{1}{251}\left(54 \times \left(-4.31 - \frac{-4.31 \times 54 + 3.26 \times 197)}{251}\right)^2 + 197 \times \left(3.26 - \frac{-4.31 \times 54 + 3.26 \times 197}{251}\right)^2\right) \\
&= 9.68
\end{aligned}$$

et la part de variabilité recherchée est donnée par $\frac{9.68}{3667.45} = 0.003$.

(c) Comme le temps de retard maximal en 2016 à Baltimore est de 160 minutes, en dénotant par x le temps de retard maximal en 2015, on a

$$160 = x - 0.07x = 0.93x$$

et donc $x = 172$.

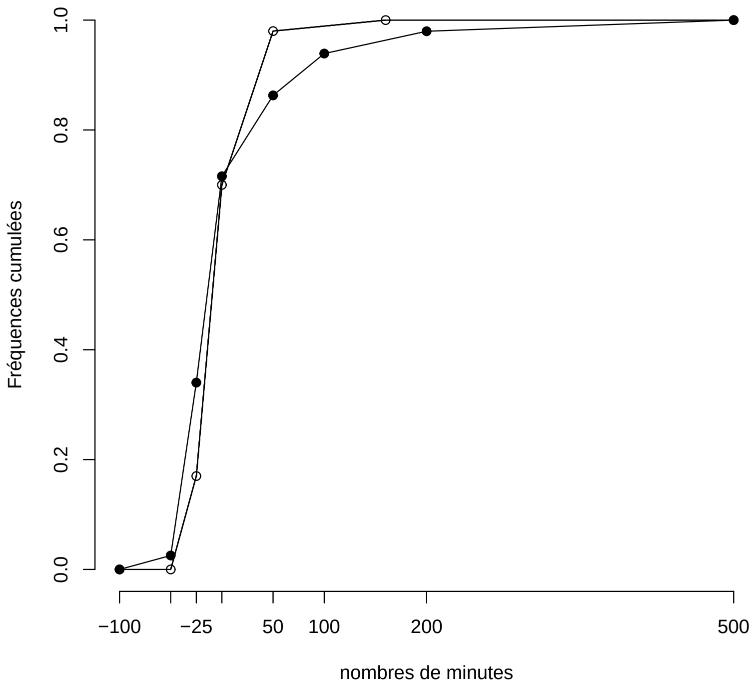

Figure 1: Ogive des fréquences cumulées

(d) Pour dessiner la boite à moustache des données de Baltimore, on a besoin des informations suivantes

- La médiane et les quartiles: $Q_1 = -20.75$, $\tilde{x} = -12.5$ et $Q_3 = 4.5$.
- Les valeurs pivots et adjacentes : on calcule tout d'abord $a_1 = Q_1 - 1.5\ EIQ_1 = -59$ et $a_2 = Q_3 + 1.5\ EIQ_1 = 42$, ce qui nous permet d'obtenir $x_{(g)} = -47$ et $x_{(d)} = 41$.
- Les valeurs extrêmes: en utilisant le diagramme en tiges et feuilles, on constate que les valeurs en dehors de l'intervalle $[x_{(g)}, x_{(d)}]$ sont données par 64, 75 et 160.

La boîte à moustaches est représentée à la Figure 2.

3. (a)
   - Dénotons par n_{ij} les effectifs bivariés ($i = 1, 2$, $j = 1, \cdots, 6$). L'effectif total n est donné dans l'énoncé (et celui-ci peut être re-calculé à partir des effectifs bivariés: $n = \sum_{i=1}^{2} \sum_{j=1}^{6} n_{ij} = 29 + 0 + 8 + 15 + 2 + 0 + 85 + 5 + 0 + 22 + 83 + 2 = 251$). La fréquence bivariée recherchée est donc donnée par $f_{13} = \frac{n_{13}}{n} = \frac{8}{251} = 0.03$.
   - Le mode recherché est la modalité correspondant au plus grand effectif bivarié n_{ij} avec $i = 2$ (i.e. dans la deuxième ligne du tableau) : Los Angeles (avec un effectif de 85).
   - Finalement, l'effectif total de la distribution des arrivées en partant de Boston est donné par $n_{2\bullet} = 85 + 5 + 0 + 22 + 83 + 2 = 197$ et donc la fréquence conditionnelle recherchée est $\frac{n_{24}}{n_{2\bullet}} = \frac{22}{197} = 0.11$.

   (b)
   - On calcule respectivement $a = 29$, $b = 0+8+15+2+0 = 25$, $c = 85$, $d = 5+0+22+83+2 = 112$ et $n = 251$.
   - Il suffit de calculer
   $$\frac{29 + 112}{251} = 0.56.$$
   - Elle vaut 0 car Boston ne fait pas partie des modalités de la variable `Destination`.

QCM

1. En mars, le nombre d'achat vaut 96% du nombre d'achats de janvier. Soit x le nombres d'achat en janvier. Le nombre d'achats en mars est donné par $(x \times 0.80) \times 1.2 = x \times 0.96$.

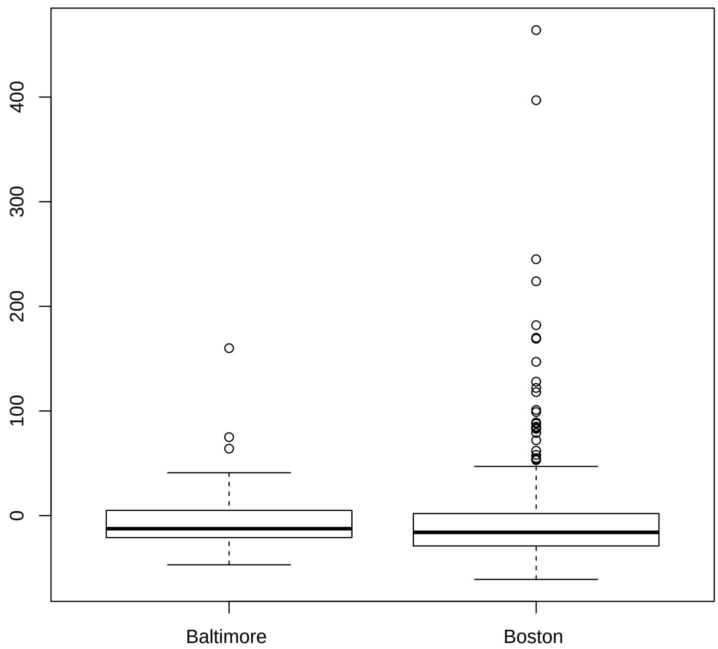

Figure 2: Boîte à moustaches des retards enregistrés à Baltimore (représentée à côté de celle basée sur les emsures obtenues à Boston)

2. - Faux. Les données étant quantitatives, l'histogramme des fréquences est ajusté et les fréquences sont donc proportionnelles à l'aire des rectangles et non à leur hauteur. Par conséquent, la classe $]7, 10]$ par exemple a clairement une fréquence plus grande.
   - Faux. Pour les mêmes raisons qu'au point précédent. Le rectangle de la classe $]1, 4]$ a clairement une aire plus grande.

3. - Vrai: Si tout les salaires augmentent de 3%, les observations restent classées dans le même ordre et, en particulier, l'observation médiane augmente de 3%.
   - Faux : Dénotons par x_i les observations de la variable "salaire" et par n le nombre d'observations. Si les salaires sont élevés au carré, la nouvelle moyenne est égale à $\frac{1}{n}\sum_{i=1}^{n} x_i^2$. Par contre, la moyenne de départ élevée au carré est donnée par $\frac{1}{n^2}\left(\sum_{i=1}^{n} x_i\right)^2$. Ces deux expressions ne sont pas égales en général.
   - Vrai: Calculons le nouveau salaire moyen
   $$\frac{1}{n}\sum_{i=1}^{n}(x_i + 100) = \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} 100 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + 100.$$
   Il s'agit bien du salaire moyen moyen augmenté de 100.

4. Il y a au minimum 88 observations entre 35 et 65. En utilisant la propriété de Tchebychev, on sait que dans l'intervalle $]50 - 15, 50 + 15[=]50 - 3s, 50 + 3s[$, il y a au minimum $100 - \frac{100}{3^2} = 88$ observations.

5. - Faux. Par exemple, la série bivariée $\{(x_i, y_i), 1 \leq i \leq n\}$ définie par $y_i = -x_i$ (avec $s_x^2 > 0$) a pour covariance $-s_x^2 < 0$.
   - Vrai. Si on translate toutes les données, les moyennes sont translatées de la même constante et vu que la covariance est définie à partir des écarts entre les observations et les moyennes, la constante s'élimine.

6. Le coefficient de détermination peut se calculer en prenant le carré du coefficient de corrélation
$$r^2 = \frac{s_{XY}^2}{s_X^2 s_Y^2} = \frac{0.9^2}{1 \times 0.95} = 0.85.$$