Statistique descriptive – Examen septembre 2019

Correctif

Bachelier en sciences informatiques

Ex. 1.(a)

- D'une part, on sait que l'effectif total vaut 74. On a donc

$$n_{23} + n_{33} = 10 \tag{1}$$

- La distribution conditionnelle de la variable `Calories` lorsque le producteur est le producteur `C` est donnée par

`Calories`	[50, 100]	]100, 110]	]110, 160]	
Effectifs	7	10	n_{33}	$17 + n_{33}$

Puisque la médiane vaut 104.5, la classe médiane est la classe $]100, 110]$. Par ailleurs, la médiane correspond à une fréquence cumulée de $\frac{1}{2}$. Ainsi, dans l'ogive des fréquences cumulées, les points $A : (100, \frac{7}{17 + n_{33}})$, $M : (104.5, \frac{1}{2})$ et $B : (110, \frac{17}{17 + n_{33}})$ sont alignés, et situés sur la droite d'équation

$$y - \frac{1}{2} = \frac{\frac{1}{2} - \frac{7}{17 + n_{33}}}{104.5 - 100}(x - 104.5)$$

Ex. 1.(a)

En injectant les cordonnées de B dans cette dernière équation et en la résolvant par rapport à n_{33}, on obtient

$$\frac{17}{17+n_{33}} - \frac{1}{2} = \frac{17+n_{33}-14}{2\times(17+n_{33})\times 4.5}(110-104.5) \Leftrightarrow \ \ldots \ \Leftrightarrow \ \boxed{n_{33}=6}$$

Finalement, de l'équation (1), on tire

$$\boxed{n_{23}=4}$$

Ex. 1.(b)

La distribution marginale de la variable `Calories` est donnée par

`Calories`	[50, 100]	]100, 110]	]110, 160]	
Effectifs	28	28	18	74

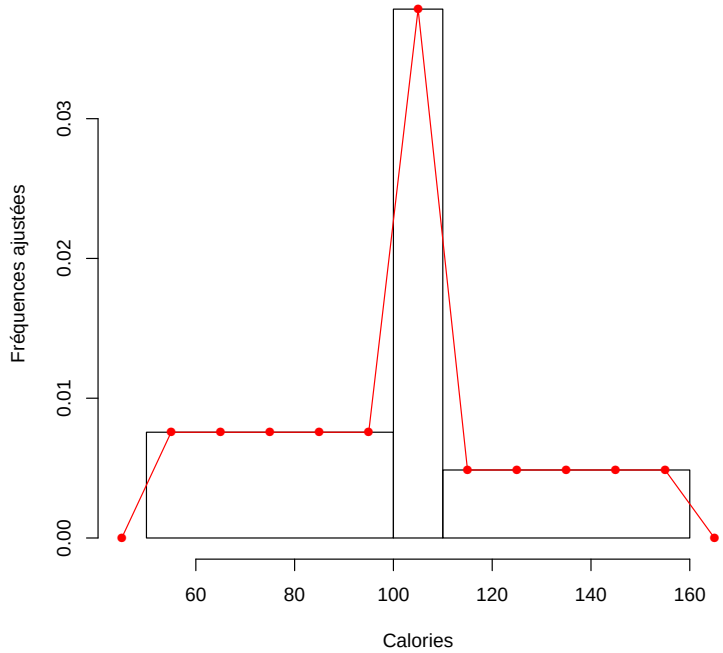
- Histogramme d'aire unitaire :

  Hauteurs des rectangles = fréquences ajustées $= \dfrac{f_j}{a_j}$

$$h_1 = \frac{28/74}{50} = 0.008;\ h_2 = \frac{28/74}{10} = 0.038;\ h_3 = \frac{18/74}{50} = 0.005$$

- Polygone des fréquences :
  Unité d'amplitude de classe $u = \text{pgcd}(50, 10, 50) = 10$

Ex. 1.(b)



La distribution de la variable `Calories` paraît plutôt symétrique autour de la classe]100, 110], bien qu'il ne soit pas aisé de conclure à ce sujet à partir de seulement 3 classes (on ne sait pas comment est distribuée le variable à l'intérieur de la première et de la dernière classe)

Ex. 2.(a)

Utilisons les notations habituelles du cours, indicées par A ou B lorsque les quantités en question ne concernent que les observations du producteur `A` ou `B`.

- Producteur `A` : L'énoncé nous fournit directement

$$n_A = 29 \; ; \; \bar{x}_A = 102 \; ; \; s_A^2 = 92^2 = 8464$$

- Producteur `B` : On a $n_B = 22$ et

$$\bar{x}_B = \frac{1}{n_B} \sum_{i=1}^{n_B} x_{i,B} = \frac{1875}{22} = 85.23$$

$$s_B^2 = \frac{1}{n_B} \sum_{i=1}^{n_B} x_{i,B}^2 - \bar{x}_B^2 = \frac{203625}{22} - \left(\frac{1875}{22} \right)^2 = 1991.99$$

- Producteurs `A` et `B` : On a $n = n_A + n_B = 51$. La moyenne globale est donnée par

$$\bar{x} = \frac{n_A \bar{x}_A + n_B \bar{x}_B}{n_A + n_B} = \frac{29 \times 102 + 1875}{51} = 94.76$$

Ex. 2.(a)

La variance globale est quant à elle donnée par

$$
\begin{aligned}
s^2 &= \text{variance dans les groupes} + \text{variance entre les groupes} \\
&= \frac{n_A s_A^2 + n_B s_B^2}{n} + \frac{n_A(\bar{x} - \bar{x}_A)^2 + n_B(\bar{x} - \bar{x}_B)^2}{n} \\
&= \frac{29 \times 8464 + 22 \times 1991.99}{51} + \frac{29 \times (94.76 - 102)^2 + 22 \times (94.76 - 85.23)^2}{51} \\
&= 5672.15 + 68.98 = 5741.13
\end{aligned}
$$

Le tableau demandé est donc finalement donné par

	Série cond. `Potassium` marque `A`	Série cond. `Potassium` marque `B`	Série `Potassium` marques `A` et `B`
Effectifs	29	22	51
Moyennes	102	85.23	94.76
Variances	8464	1991.99	5741.13

Ex. 2.(b) et 2.(c)

- La part de la variance totale expliquée par la variabilité entre les groupes est donnée par

$$\frac{68.98}{5741.13} = 0.01$$

  Cela signifie donc que seul 1% de la variabilité du contenu en potassium provient de la variabilité entre les groupes et, dès lors, que 99% provient de la variabilité dans les groupes. Le producteur B disait donc vrai.

- On a

$$(n_A + n_B - 1)\frac{\text{Variance entre les groupes}}{\text{Variance dans les groupes}} = 50\frac{68.98}{5672.13} = 0.61$$

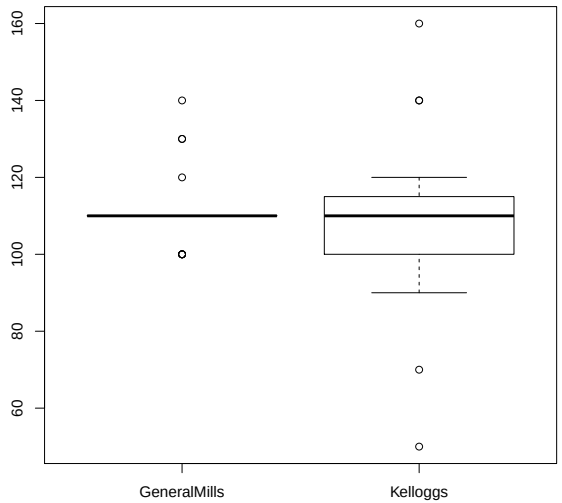
  L'inégalité n'est donc pas vérifiée et les deux moyennes ne sont donc pas significativement différentes. Ceci appuie donc l'argument du producteur B.

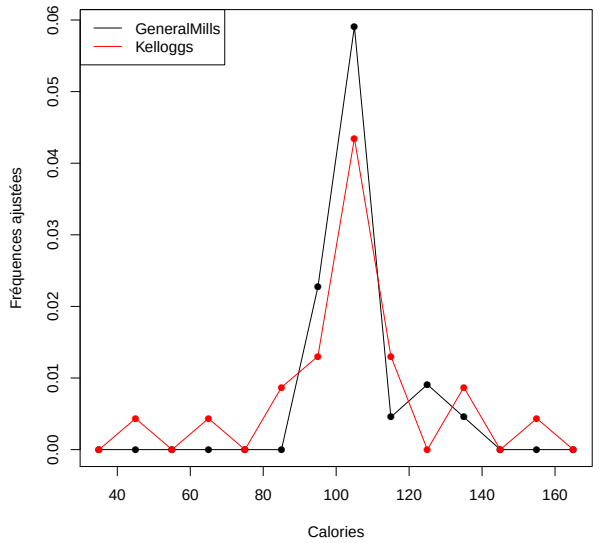
Analyse de données

Ces transparents ainsi que la vidéo de présentation associée se concentrent sur la présentation et l'interprétation des résultats.

Un code reprenant les commandes du logiciel `R` utilisées pour obtenir les différents résultats est fourni à part.

Analyse de données – Question 1.(a)





- Les boîtes à moustaches permettent clairement de détecter une différence de dispersion dans les deux populations. Les observations de la marque `GeneralMills` sont presque toutes concentrées en la valeur 110 alors que celles de la marque `Kelloggs` sont plus dispersées.
- Les tendances centrales sont similaires dans les groupes, comme le montrent les deux graphiques.
- La distribution de la marque `GeneralMills` présente un étalement sur la droite alors que la distribution de la marque `Kelloggs` semble plus symétrique (avec toutefois un petit étalement sur la gauche).

Analyse de données – Question 1.(b)

	`GeneralMills`		`Kelloggs`
Moyenne	111.36	$\approx$	108.70
Médiane	110	$=$	110
Ecart-Type	10.37	$<$	22.22
EIQ	0	$<$	15
Dissymétrie (Fisher)	1.30		-0.32

Les paramètres statistiques donnés ci-dessus confirment les observations graphiques :

- Tendance centrale : Moyennes et médianes similaires ;
- Dispersion : écart-type plus important pour la marque `Kelloggs` et écart interquartile nul pour la marque `GeneralMills` (les 50% d'observations centrales sont égales) ;
- Symétrie : coefficient de dissymétrie positif pour la marque `GeneralMills` (étalement sur la droite) et légèrement négatif pour la marque `Kelloggs` (léger étalement sur la gauche).

Analyse de données – Question 1.(c)

La quantité $\left|\frac{\bar{x}_K - \bar{x}_G}{s}\right|$ vaut 0.52 et n'est donc pas supérieure à 2.036.

On ne peut donc pas conclure à une différence significative entre les deux types de céréales. On avait déjà observé précédemment, via les boîtes à moustaches et les polygones des fréquences ainsi que via les valeurs des moyennes et médianes, que les tendances centrales des deux marques étaient assez proches.

Analyse de données – Questions 2.(a) et 2.(b)

- Corrélations entre la variable `Potassium` et les autres variables :
  - La corrélation la plus faible vaut 0.001 et est celle calculée avec la variable `Sucres`.
  - La corrélation la plus forte vaut 0.912 et est celle calculée avec la variable `Fibre`.
- La droite de régression permettant d'expliquer la variable `Potassium` en fonction de la variable `Fibre` et estimée par la technique des moindres carrés a pour équation

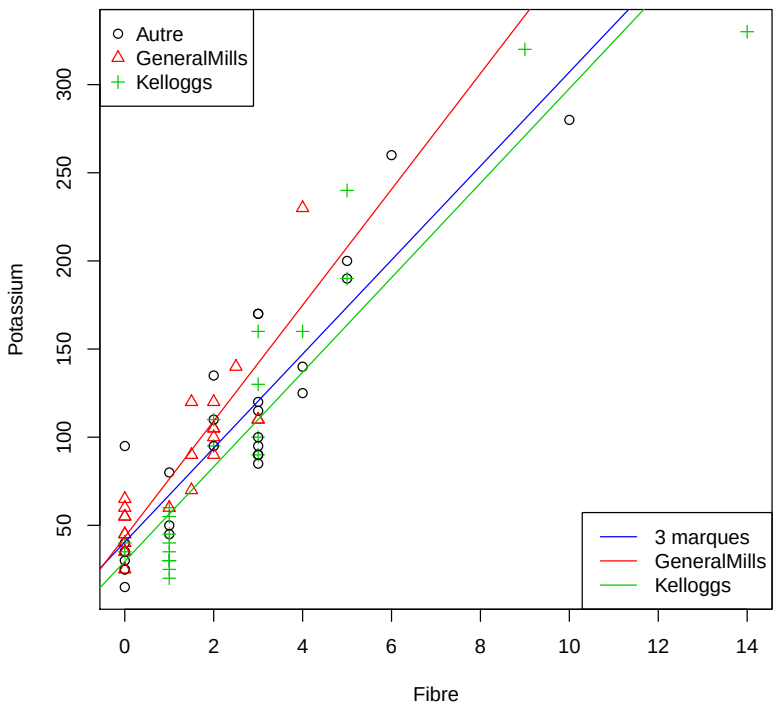
$$\texttt{Potassium} = 40.51 + 26.66\,\texttt{Fibre}.$$

  - Le résidu des céréales `Smacks` vaut -27.17 (ce qui signifie que l'on surestime la vraie valeur).
  - La valeur ajustée des céréales `Golden_Crisp` vaut 40.51.

Analyse de données – Questions 2.(c) et 2.(d)

- La qualité de l'ajustement peut être mesurée à l'aide du coefficient de détermination, qui vaut 0.83.

  Cela signifie que 83% de la variance de la variable `Potassium` est expliquée par la régression, qui est donc d'une bonne qualité.
- Les corrélations entre les deux variables valent 0.89 dans le sous-ensemble des céréales `GeneralMills` et 0.93 dans le sous-ensemble des céréales `Kelloggs`.



La droite de régression globale et la droite de régression des céréales `Kelloggs` sont assez similaires alors que la droite de régression des céréales `GeneralMills` se distingue légèrement. Les deux premières droites semblent être attirées par les quelques observations du coin supérieur droit. Puisque ces observations ne concernent pas la marque `GeneralMills`, la droite associée à ces céréales ne l'est pas.