DM1: Correction

Thibault Lagache

1. Pour ces deux distributions, quelle est la population? Quelle est la variable? De quel type de variable s'agit-il?
   La population est soit les 6000 **clients du mercredi, soit les 6000 clients du samedi. La variable est le montant du ticket de caisse. C'est une variable quantitative de type ratio.**

2. Tracer les histogrammes des deux échantillons (en fréquence relative) avec les cellules pré-définies par l'énoncé.
   On commence par calculer les hauteurs de l'histogramme pour chaque cellule, pour le mercredi:

	montant du ticket en euros	proportion des tickets	hauteur de l'histogramme
i=1	de 5 à moins de 30	0.09	0.09/25=0.0036
i=2	de 30 à moins de 60	0.12	0.12/30=0.004
i=3	de 60 à moins de 100	0.22	0.22/40=0.0055
i=4	de 100 à moins de 150	0.25	0.25/50=0.005
i=5	de 150 à moins de 200	0.19	0.19/50=0.0038
i=6	de 200 à 300	0.13	0.0013

et le samedi..

	montant du ticket en euros	proportion des tickets	hauteur de l'histogramme
i=1	de 5 à moins de 30	0.20	0.20/25=0.008
i=2	de 30 à moins de 60	0.14	0.0047
i=3	de 60 à moins de 100	0.16	0.004
i=4	de 100 à moins de 150	0.19	0.0038
i=5	de 150 à moins de 200	0.13	0.0026
i=6	de 200 à 300	0.18	0.0018

On suppose à présent que les montants des tickets dans chaque intervalle de prix sont uniformément répartis.

1. En déduire, pour chaque jour, la moyenne, l'écart-type et les différents quartiles. Discuter.

	fréquence relative	frequence relative cumulée	moyenne $\bar{x}_i$
i=1	0.09	0.09	(30+5)/2=17.5
i=2	0.12	0.21	45
i=3	0.22	0.43	80
i=4	0.25	0.68	125
i=5	0.19	0.87	175
i=6	0.13	1	250

et le samedi:

	fréquence relative	frequence relative cumulée	moyenne $\bar{x}_i$
i=1	0.20	0.20	17.5
i=2	0.14	0.34	45
i=3	0.16	0.5	80
i=4	0.19	0.69	125
i=5	0.13	0.82	175
i=6	0.18	1	250

La moyenne $\bar{x}$ est la moyenne pondérée $\bar{x} = \sum_{i=1}^{6} f_i \bar{x}_i$, où f_i est la fréquence relative de chaque cellule et $\bar{x}_i$ la moyenne de chaque cellule. On a donc $\bar{x} = 17.5 * 0.09 + 45 *$

$0.12 + \ldots + 0.13 * 250 = 121.6$. **Vu les fréquences cumulées, le premier quartile $q_{0.25}$ se trouve dans la cellule $[60; 100[$ et comme les montants des tickets de caisse sont supposés uniformément répartis, on trouve:** $q_{0.25} = 60 + 0.04/0.22 * (100 - 60) = 67.3$. **De même, on a** $q_{0.5} = m = 100 + 0.07/0.25 * (150 - 100) = 114$ **et** $q_{0.75} = 150 + 0.07/0.19 * (200 - 150) = 168.4$. **Enfin le calcul de l'écart-type $\sigma = \sqrt{var}$ est plus fastidieux, On a en effet la variance var qui est égale à (la cellule x_i et x_{i+1} sont les bornes de la cellule numéro i):**

$$var = \sum_{i=1}^{6} \frac{f_i}{x_{i+1} - x_i} \int_{x_i}^{x_{i+1}} (x - \bar{x})^2 dx = \sum_{i=1}^{6} \frac{f_i}{x_{i+1} - x_i} \int_{x_i}^{x_{i+1}} x^2 dx - \bar{x}^2 \quad (1)$$

$$= \sum_{i=1}^{6} \frac{f_i}{x_{i+1} - x_i} \frac{x_{i+1}^3 - x_i^3}{3} - \bar{x}^2 = \sum_{i=1}^{6} \frac{f_i}{3} \left(x_i^2 + x_i x_{i+1} + x_{i+1}^2\right) - \bar{x}^2 \quad (2)$$

ce qui nous donne numériquement:

$$var = 0.09/3 * (5^2 + 5 * 30 + 30^2) + \ldots + 0.13/3 * (200^2 + 200 * 300 + 300^2) - 121.6^2 = 4985, \quad (3)$$

et l'écart type est égal à $\sigma = \sqrt(var) = 70$.
Les calculs pour la journée sont similaires et nous donnent notamment une moyenne $\bar{x} = 114.1$ et un écart-type $\sigma = 83$. On remarque donc que les montants de tickets de caisse sont plus dispersés le samedi, notamment à cause du fait que l'on a une plus grande proportion de petits (montant entre 5 et 30 euros) et de gros achats (supérieurs à 200 euros).

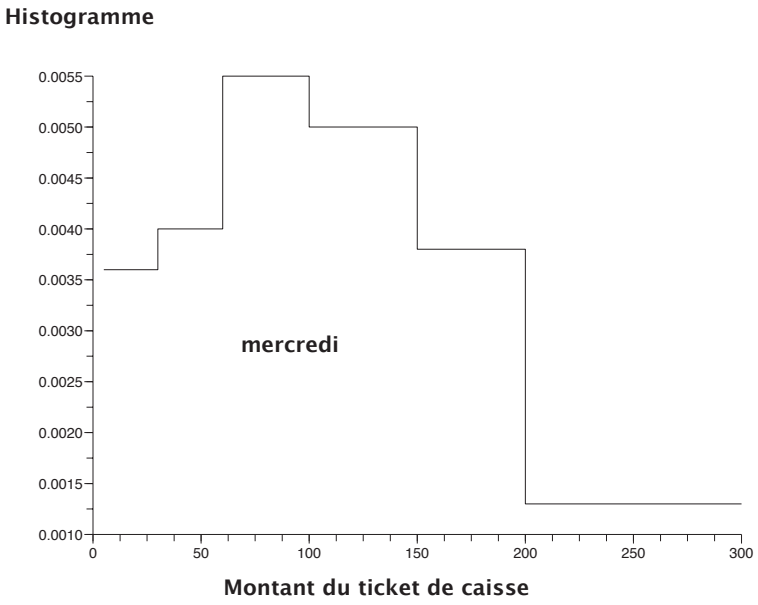
2. Tracer les fonctions de répartition des 2 échantillons et représenter graphiquement la distance de Kolmogorov-Smirnov.
La distance maximale entre les deux fonctions de répartition est atteinte en $x = 60$ et vaut $d_{KS} = 0.13$, ce qui est assez important..

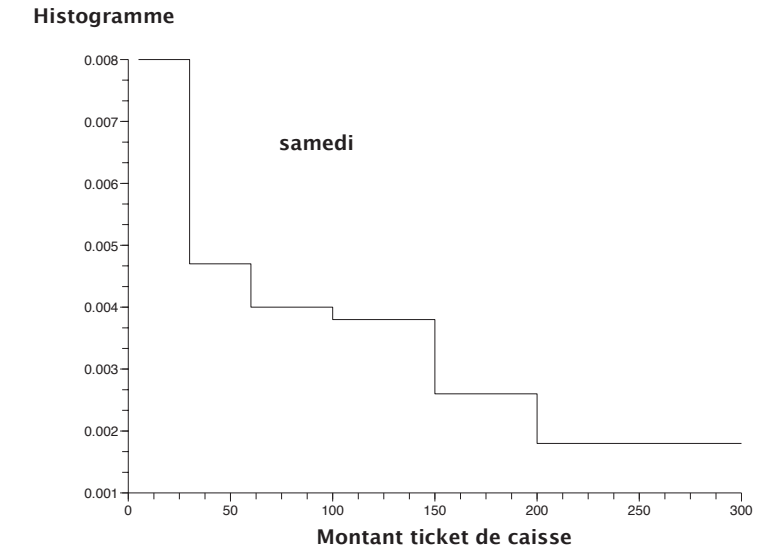
Pour un montant de caddie $5 \leq m \leq 300$, on modélise le temps t nécessaire pour passer en caisse par $t = t_0 + \alpha m + e^{\frac{m - m_0}{\beta}}$ avec $t_0 = 5$ minutes, $\alpha = \frac{1}{30}$, $m_0 = 150$ et $\beta = 150/ln(10)$.

1. Selon vous, quelle est l'origine de chacun des trois termes t_0, αm et $e^{\frac{m - m_0}{\beta}}$.
t_0 **représente le temps incopressible passé en caisse, indépendamment du montant de courses (temps pour payer..),** αm **est proportionnel au montant de courses et représente par exemple le temps pour lire le code barre de chaque article, les ranger dans les sacs plastiques... Enfin,** $e^{\frac{m - m_0}{\beta}}$ **indique que le temps peut exponentiellement exploser avec le montant de courses (nécessité de prendre 2 caddies, tapis trop court pour déposer tous les articles..**

2. Calculer, pour chaque intervalle de montant, le temps moyen $\bar{t}_i$, $1 \leq i \leq 6$ (on supposera encore que le montant des caddies est uniformément réparti dans chaque intervalle). **Il faut calculer ici, pour chaque cellule:** $\bar{t}_i = \frac{f_i}{x_{i+1} - x_i} \int_{x_i}^{x_{i+1}} t_0 + \alpha m + e^{\frac{m - m_0}{\beta}} dm$**...Le calcul est un peu long (désolé!) mais sans difficultés!**

3. En déduire, pour chaque jour le temps de caisse moyen total $\bar{t} = \sum_{i=1}^{6} \bar{t}_i$ et le nombre de caisse $C = \bar{t}/400$ nécessaire. Conclure.


Histogramme
0.0055
0.0050
0.0045
0.0040
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0
50
100
150
200
250
300
mercredi
Montant du ticket de caisse


Histogramme
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
50
100
150
200
250
300
samedi
Montant ticket de caisse

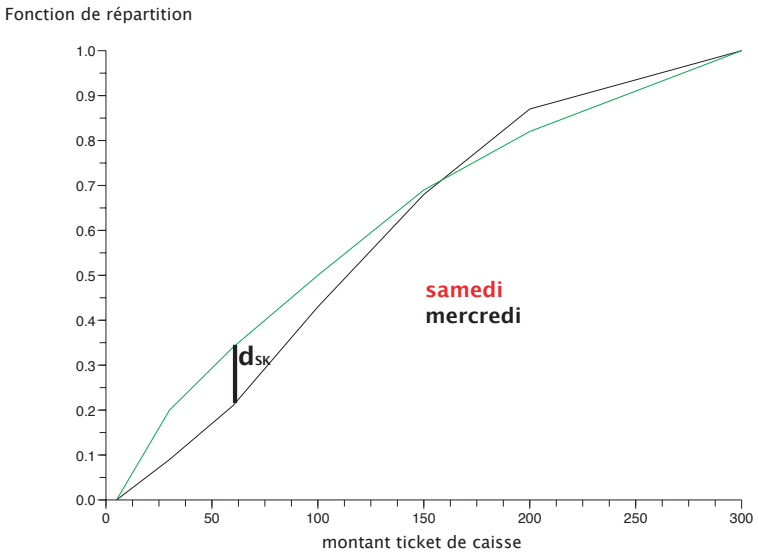

Fonction de répartition
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
50
100
150
200
250
300
samedi
mercredi
dSK
montant ticket de caisse

FIG. 1: