Université Claude Bernard - Lyon 1 **Semestre automne 2022-2023**

L2 - Informatique UE : Statistiques pour l'Informatique

Feuille TD 1

Exercice 1.6 Un atelier réalise le séchage de boues d'origine industrielle. Il obtient à la fin du processus des déchets. On a observé les poids suivants de déchets après le traitement de 100 kg de boues :

$$4,7 \quad 4,3 \quad 4,5 \quad 4,9 \quad 4,2 \quad 4,7 \quad 4,0 \quad 4,2 \quad 5,0 \quad 3,9 \quad 4,6 \quad 4,6$$

$$4,8 \quad 4,4 \quad 4,2 \quad 4,6 \quad 4,3 \quad 4,9 \quad 4,0 \quad 4,5 \quad 4,1 \quad 4,4 \quad 4,3 \quad 4,3$$

Notons x cette série statistique.

a) Opérer le dénombrement des différentes modalités du caractère et construire le tableau des effectifs, fréquences, fréquences cumulées.

L'effectif total est 24

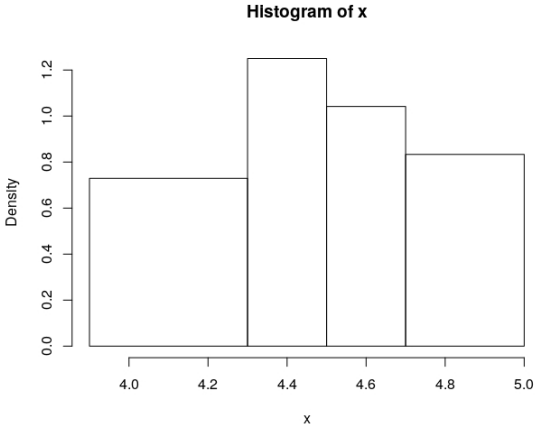
	3.9	4	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5
$effectif$	1	2	1	3	4	2	2	3	2	1	2	1
$fréquence$	$\frac{1}{24}$	$\frac{1}{12}$	$\frac{1}{24}$	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{24}$	$\frac{1}{12}$	$\frac{1}{24}$
$fréquence\ cumulée$	$\frac{1}{24}$	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{7}{24}$	$\frac{11}{24}$	$\frac{13}{24}$	$\frac{15}{24}$	$\frac{18}{24}$	$\frac{20}{24}$	$\frac{21}{24}$	$\frac{23}{24}$	1

b) Tracer un histogramme. On choisit les classes pour que l'effectif de chaque classe soit au moins 5

On prend comme subdivision : $a_0 = 3.9, a_1 = 4.3, a_2 = 4.5, a_3 = 4.7, a_4 = 5$.

On calcule $N([a_i, a_{i+1}[)\ i < n-1$ et $N([a_{n-1}, a_n])$. puis $h_i = \frac{N([a_i, a_{i+1}[)}{N(a_i - a_{i-1})}$

	$[3.9, 4.3[$	$[4.3, 4.5[$	$[4.5, 4.7[$	$[4.7, 5]$
$effectif$	7	6	5	6
$hauteur$	$\frac{7}{9.6} \simeq 0.72$	$\frac{6}{4.8} \simeq 1.25$	$\frac{5}{4.8} \simeq 1.04$	$\frac{6}{7.2} \simeq 0.83$



c) Déterminer la médiane et les quartiles. Tracer le diagramme à moustache (boxplot).

On utilise les fréquences cumulées pour trouver les quartiles.

(a) La première fréquence cumulée au dessus de 1/4 est 7/24, la valeur correspondante est $Q_1(x) = 4.2$.

(b) La première fréquence cumulée au dessus de 1/2 est 13/24, donc $Q_2(x) = 4.4$ Comme l'échantillon est pair de longueur $N = 2l$, cette valeur coïncide avec x^*_{l+1} de la définition de la médiane, il se pourrait que la médiane soit différente. Mais ici $x^*_l = 4.4$ car $(13-1)/24 \geq 1/2$. $x^*_{12} = x^*_{13} = 4.4$ donc la médiane est 4.4.

(c) La première fréquence cumulée au dessus de 3/4 est 18/24, donc $Q_3(x) = 4.6$

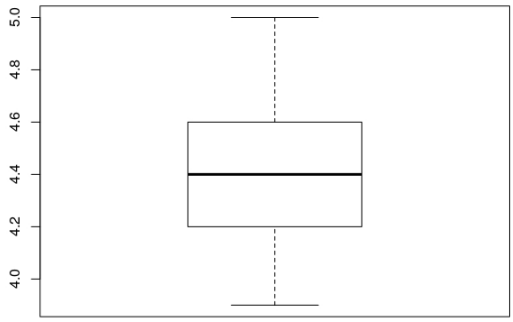
Pour tracer le diagramme à moustache, on calcule $IC = 4.6 - 4.2 = 0.4$.

Calcul de la moustache haute :

$Q_3(x) + 1.5 * IC = 4.6 + 0.6 = 5.2$ est au dessus du max de l'échantillon 5 donc $M_H = 5$ (la plus grande valeur de l'échantillon en dessous de 5.2

Calcul de la moustache basse : $Q_1(x) - 1.5 * IC = 4.2 - 0.6 = 3.6 < 3.9$ donc de même $M_B = 3.9$

Il n'y a pas de valeurs extrèmes :

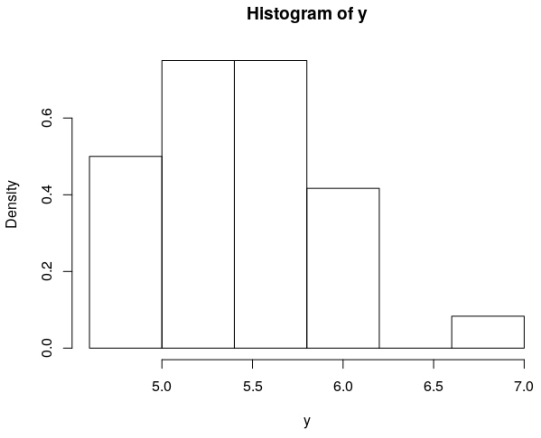


Exercice 1.7 On prélève $n = 30$ échantillons de pluies provenant du sud de la Pologne. On mesure le pH de ces échantillons et on note cette série statistique $(y_i)_{1 \leq i \leq 30}$:

4.60,	4.79,	4.81,	4.82,	4.86,	4.89,	5.03,	5.06,	5.10,	5.14
5.17,	5.18,	5.28,	5.28,	5.32,	5.44,	5.45,	5.55,	5.62,	5.63
5.64,	5.70,	5.77,	5.79,	5.81,	5.82,	5.83,	5.85,	5.97,	6.92

a) Faire un histogramme de ces données. On choisira un nombre de classes égal à l'entier supérieur à $1 + \log(n)/\log(2) = 5.907$. On prend 6 classes. On prend des classes égales entre 4.6 et 7 donc de pas 0.4
On obtient le tableau des effectifs et hauteurs $h([a, b[) = \frac{N([a,b[)}{n*0.4}$:

	$[4.6, 5[$	$[5, 5.4[$	$[5.4, 5.8[$	$[5.8, 6.2[$	$[6.2, 6.6[$	$[6.6, 7[$
$effectif$	6	9	9	5	0	1
$hauteur$	$\frac{6}{12} = 0.5$	$\frac{9}{4.8} \simeq 0.75$	$\frac{9}{4.8} \simeq 0.75$	$\frac{5}{12} \simeq 0.417$	0	$\frac{1}{12} \simeq 0.083$



b) Calculer la médiane et les quartiles.

On remet l'échantillon dans l'ordre croissant :

$$y^* = (4.6, 4.79, 4.81, 4.82, 4.86, 4.89, 5.03, 5.06, 5.1, 5.14, 5.17, 5.18, 5.28, 5.28, 5.32,$$
$$5.44, 5.45, 5.55, 5.62, 5.63, 5.64, 5.7, 5.77, 5.79, 5.81, 5.82, 5.83, 5.85, 5.97, 6.92)$$

L'effectif total est $n = 30 = 2l$. $x^*_{15} = 5.32, x^*_{16} = 5.44$ donc la médiane est $mediane(x) = \frac{5.32+5.44}{2} = 5.38$
Le deuxième quartile est $Q_2(x) = x^*_{15} = 5.32$

$30/4 = 7.5$ donc l'entier suivant est 8, donc le premier quartile est $Q_1(x) = x^*_8 = 5.06$.

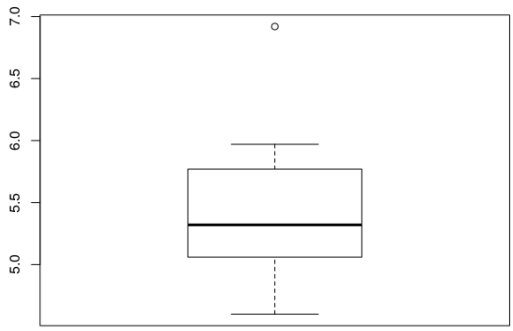
$3 * 30/4 = 22.5$ donc l'entier suivant est 23, donc le troisième quartile est $Q_3(x) = x^*_{23} = 5.77$

c) Tracer le diagramme à moustache (boxplot).

On calcule l'écart interquartile $IC = 5.77 - 5.06 = .71$

Calcul de la moustache haute : $Q_3(x) + 1.5 * IC = 5.77 + 1.055 = 6.835 > 5.97$ donc $M_H(x) = 5.97$. L'unisque valeur au dessus 6.92 est une valeur extrême.

Calcul de la moustache basse :$Q_1(x) - 1.5 * IC = 5.06 - 1.055 = 4.005 < 4.6$ le minimum de la distribution donc $M_B(x) = 4.6$.



Exercice 1.8 Soit x la variable qui décrit les caractéristiques numériques d'un échantillon de taille L.

1. La déviation de la caractéristique x_i de l'individu i est $d_i := x_i - \bar{x}$. Montrons que $\bar{d} = 0$. Ils suffit de noter que par linéarité : $\bar{d} = m(x - \bar{x}) = m(x) - m(\bar{x}) = m(x) - \bar{x} = 0$.
2. Montrons que la variance non-biaisée $var(x)$ de l'échantillon peut être obtenue par les formules suivantes

$$var(x) := \frac{1}{L-1}\sum_{i=1}^{L} d_i^2 = \frac{1}{L-1}\left(\sum_{i=1}^{L} x_i^2 - \frac{(\sum_{i=1}^{L} x_i)^2}{L}\right) = \frac{L}{L-1}\big(\overline{x^2} - (\overline{x})^2\big).$$

   On a vu en cours que $var(x) = \frac{L}{L-1}V(x)$ et on a vu $V(x) = m(x^2) - m(x)^2$ ceci donne la dernière égalité. En remplaçant les moyennes par leurs valeurs $m(x^2) = \frac{1}{L}\sum_{i=1}^{L} x_i^2$ et $m(x)^2 = \frac{1}{L^2}(\sum_{i=1}^{L} x_i)^2$ on obtient l'égalité du milieu.
3. Par l'homogénéité vue en cours $\frac{var(10x)}{var(x)} = 10^2 = 100$?

Il arrive parfois que si on considère un jeu de données globalement, on observe une certaine tendance, alors que si on sépare les données en plusieurs catégories, on observe une tendance différente - apparemment contradictoire. Ce phénomène s'appelle « **Paradoxe de Simpson** ». L'exercice suivant illustre comment cela peut se produire et pourquoi il faut être prudent avant de tirer des conclusions des données, et la nécessité de chercher des variables latentes qui puissent expliquer la contradiction apparente.

Exercice 1.9

Le journal local a examiné les deux hôpitaux de la ville, et a constaté qu'à l'hôpital Cochin, 79% des patients des six derniers mois ont survécu, tandis qu'à l'hôpital Conté 90 % des patients ont survécu. Le tableau ci-dessous résume les résultats.

	survie	décès	Total	taux de survie (en %)
Cochin	790	210	1000	79.0
Conté	900	100	1000	90.0

Dans une étude plus approfondie, il a été observé que les patients étaient catégorisés lors de l'admission comme étant en condition raisonnable (ou meilleure) ou en condition médiocre (ou pire). Quand les taux de survie ont été examinés pour ces groupes, les tableaux suivants ont été obtenus :

Patients admis avec condition raisonnable ou meilleure :

	survie	décès	Total	taux de survie
Cochin	580	10	590	0.98
Conté	860	30	890	0.96

Patients admis avec une médiocre condition ou pire :

	survie	décès	Total	taux de survie
Cochin	210	200	410	0.51
Conté	40	70	110	0.36

1. Remplissez les quatre cases dans la dernière colonne dans les deux tableaux ci-dessus avec les pourcentages corrects.
2. Comparez les pourcentages dans le premier tableau avec ceux des deux tableaux suivants. $0.98>0.96$ et $0.51>0.36$ donc Cochin est meilleur dans les deux sous-groupes de malade. Mais globalement, son taux de survie est moins bon de $0.79<0.9$. Est-ce que vous observez quelque chose d'étrange ? Oui, les inégalités sont inversés entre les deux sous-groupes et le groupe total.
3. Quel hôpital choisiriez-vous, et pourquoi ? Quelle est la variable latente ? On choisirait Cochin, car il a un meilleur taux de survie dans toute catégorie. La variable latente est la gravité de la maladie des patients admis (Cochin reçoit plus de malade gravement malade 41% contre 11%. Probablement car, étant meilleur, les patients les plus gravement malades préfèrent aller au meilleur hopital qui a un bien meilleur taux de survie pour eux). Ce paradoxe sera expliqué par la formule des probabilités totales :

$$0.79 \simeq 0.98 * 0.59 + 0.51 * 0.41$$

$$0.90 \simeq 0.96 * 0.89 + 0.36 * 0.11$$