TD 1 - Statistiques descriptives

1. – Diagramme en tige et feuilles (plus facile à construire après avoir trié les observations par ordre croissant) :

```
 5|56
 6|002468
 7|03478
 8|03
 9|
10|3
```

Cette représentation met bien en évidence l'allure générale de la distribution, et la présence d'une valeur atypique (103).

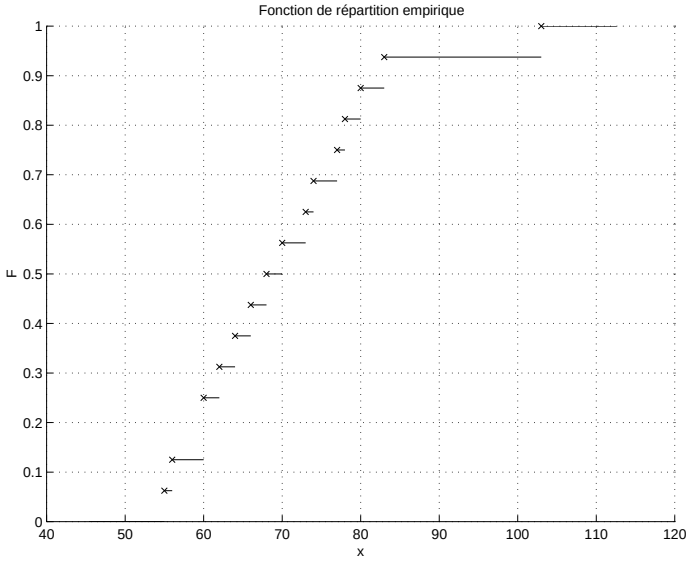

FIGURE 1 – Fonction de répartition empirique

– Résumés numériques : $\overline{x} = 70.56$, $M_1 = 69.36$, $s^{*2} = 148.80$, $s^* = 12.20$, $\widehat{f}_{0.5} = 68$, $\widehat{f}_{0.25} = 60$, $\widehat{f}_{0.75} = 77$, $W = 103 - 55 = 48$, $H = 77 - 60 = 17$.
On remarque que la moyenne empirique est supérieure à la médiane, ce qui traduit une légère dissymétrie due à la valeur atypique 103.

2. (a) Tableau de fréquences :

c_k	a_k	a_{k+1}	n_k	f_k	F_k
1750	1500	2000	510 10^3	0.048	0.048
2250	2000	2500	1290 10^3	0.121	0.169
2750	2500	3000	4070 10^3	0.381	0.549
3500	3000	4000	4094 10^3	0.383	0.933
5000	4000	6000	720 10^3	0.067	1

(b) Histogramme : dans un histogramme les *aires* des rectangles associés à chaque classe sont proportionnelles aux fréquences. Cela revient à dire que les hauteurs des rectangles sont proportionnelles à $f_k/(a_{k+1} - a_k^)$. En notant h_k la hauteur du rectangle définissant la classe k, et en prenant un coefficient de proportionalité égal à 1 (de cette façon, l'aire de l'histogramme sera égale à 1), on a :

c_k	a_k	a_{k+1}	h_k ($\times 10^{-3}$)
1750	1500	2000	0.0955
2250	2000	2500	0.2415
2750	2500	3000	0.7619
3500	3000	4000	0.3832
5000	4000	6000	0.0337

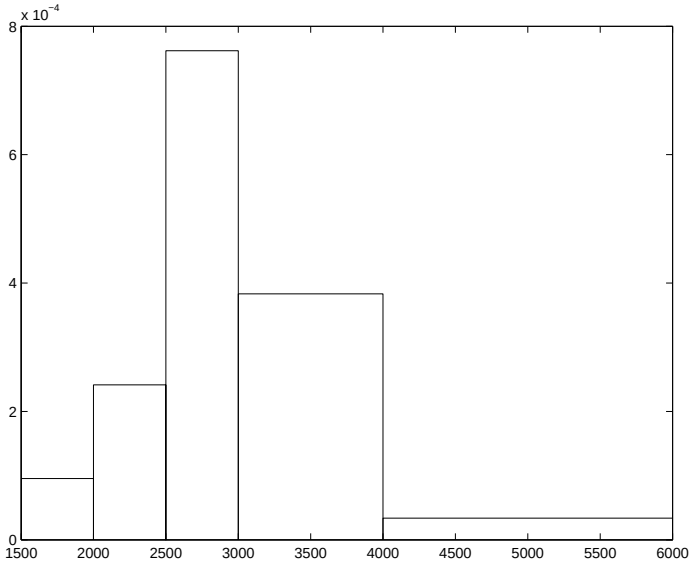

FIGURE 2 – Histogramme

(c) Si on remplace chaque valeur par le centre de sa classe, on obtient les approximations suivantes : $\overline{x} = 3080.9$, $s = 724.65$. La classe modale est la classe de plus grande densité, soit la classe $[2500, 3000]$.

3. (a) La variable « situation familiale » est une variable qualitative nominale. Les autres variables sont numériques.

Pour la variable « situation familiale », on ne calcule pas les fréquences cumulées qui n'ont aucun sens (l'ordre des modalités est arbitraire). Pour les variables quantitatives, il faut faire un découpage en classes (pas plus de 3 ou 4 étant donnée la faible taille de l'échantillon).

modalités	n_k	f_k
M	5	0.333
W	1	0.067
B	2	0.133
BM	2	0.133
Y	3	0.2
C	2	0.133

TABLE 1 – Tableau de fréquences pour la variable « situation familiale ».

classes	n_k	f_k	F_k
0-200	4	0.267	0.267
201-300	4	0.267	0.533
301-350	4	0.267	0.8
351-	3	0.2	1

TABLE 2 – Tableau de fréquences pour la variable « nuits ».

(b) Boîtes à moustaches : il faut d'abord calculer les quartiles pour chacune des deux distributions.

Javelot : $\widehat{f}_{0.25} = 2$, $\widehat{f}_{0.5} = 8.1$, $\widehat{f}_{0.75} = 16.2$.

Hameçon : $\widehat{f}_{0.25} = 2.5$, $\widehat{f}_{0.5} = 8$, $\widehat{f}_{0.75} = 14.9$. On remarque que, pour la variable « javelot », il y a deux points extrêmes, supérieurs à $\widehat{f}_{0.75} + 1.5(\widehat{f}_{0.75} - \widehat{f}_{0.25})$.

On ne remarque pas de différence notable entre les deux distributions (à l'exception des deux points extrêmes).

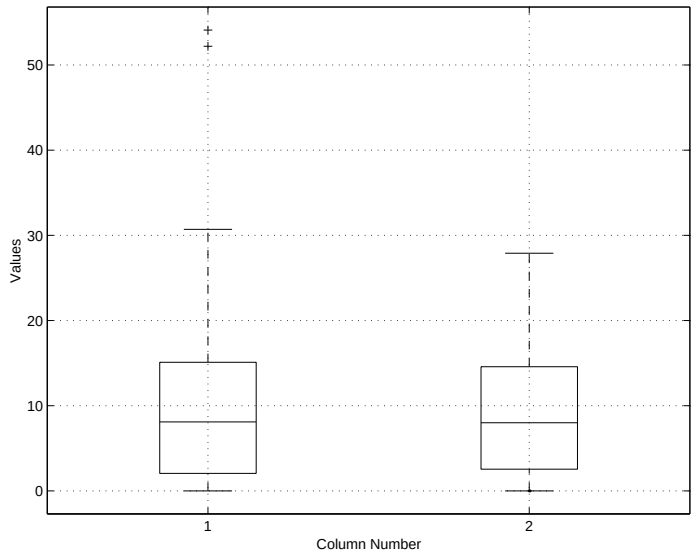

FIGURE 3 – Boîtes à moustaches pour les variables « javelot » (gauche) et « hameçon » (droite)