Solutions Fiche 1 et Exercice du Cours2 - diapo n°44

Exercice 1.2

Les données proposées ici sont un échantillon de données brutes de taille $n = 12$, c'est à dire qu'on donne la suite $(x_1, ..., x_{12})$. Dans le morceau de script ci-dessous on saisit les données et on les range dans un objet nommé `x`. Pour savoir si on regarde cet échantillon comme celui d'une variable discrète ou continue on dresse son tableau en effectifs. Si on observe beaucoup de répétitions de valeurs dans l'échantillon on considèrera que les données sont discrètes dans le sens où la variable aléatoire que l'on choisira pour modéliser sera une variable aléatoire discrète (pour laquelle $\mathcal{X}$ sera fini ou infinini dénombrable), et si au contraire il y a peu de répétitions observées dans l'échantillon de données la variable aléatoire qui servira de modèle sera continue (c. à d. que $\mathcal{X} \subset R$).

```
x<-c(101,104,102,99,101,98,97,104,100,96,103,101)
# saisie des données à mettre dans une liste de la calculette

table(x)
```

```
## x
##  96  97  98  99 100 101 102 103 104
##   1   1   1   1   1   3   1   1   2
```

```
# calcul des effectifs observés par valeurs différentes rencontrées (modalités)
```

Ici on traitera donc la variable observée comme une variable continue.

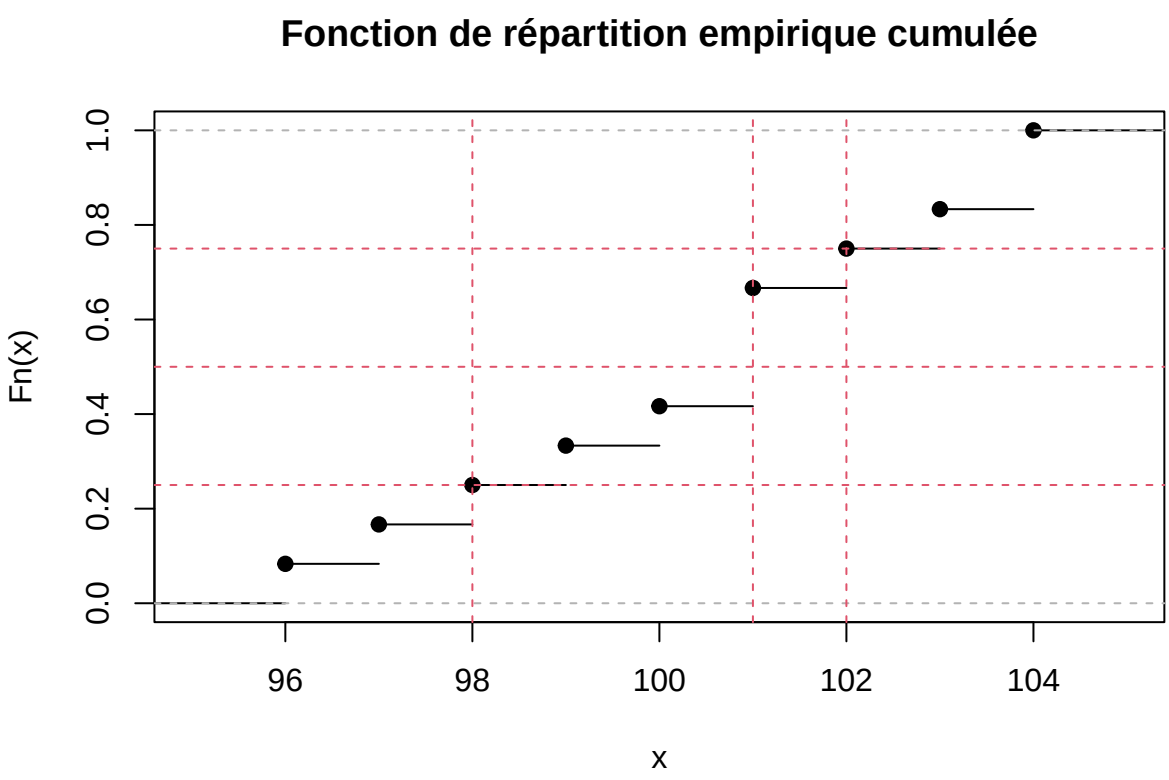
question 1 :

A l'exception des modalités 101, resp. 104 pour lesquelles on observe 3, resp. 2 répétitions les autres valeurs rencontrées dans l'échantillons ne le sont qu'une seule fois. L'échantillon étant de taille 12 avec 10 modalités différentes on considèrera que la variable observée X (poids d'un fichier donné en ko) est une variable quantitative continue. Si on avait eu moins de modalités (par ex deux ou trois) et plus de répétitions (des effectifs entre 1 et 6 pour se fixer les idées) on regarderait cette variable comme une variable discrète.

question 2 :

Pour calculer un quantile empirique d'ordre α on peut utiliser la réprésentation de la fonction de répartition empirique cumulée et repérer sur ce graphe la première modalité (elles sont classées dans l'ordre croissant) à partir de laquelle la fonction dépasse (ou est égale) à α. Ici on s'intéresse aux quartiles, c'est à dire aux quantiles empiriques d'ordre 25%, 50% (la médiane) et 75% notés respectivement q_1, q_2 et q_3.

```
plot(ecdf(x), main="Fonction de répartition empirique cumulée")
abline(h=c(0.25,0.5,0.75),lty=2,col=2)
abline(v=c(98,101,102), lty=2,col=2)
```



On lit en abscisse $q_1 = 98$, $q_2 = 101$ et $q_3 = 102$.

Remarque : selon les calculatrices où les fonctions R utilisées, les quantiles empiriques peuvent ne pas être définis comme une valeur de l'échantillon. Par exemple si il y a une nombre pair de valeurs dans l'échantillon les calculatrices prendront comme médiane le milieu de $[x_{n/2}, x_{n/2+1}]$. La fonction `quantile` de R propose sept définitions différentes des quantiles empiriques (caractérisée par l'option `type`). On convient dans ce cours d'utiliser la définition du quantile empirique d'ordre α comme la plus petite modalité à partir de laquelle la fonction de répartition empirique cumulée dépasse α (au sens large), qui correspond au choix `type=1` dans la fonction `quantile` de R et est la définition de quantile empirique donnée en cours et dans le poly.

Si on ne veut pas utiliser la fonction de répartition empirique cumulée on peut utiliser le menu de la calculatrice qui retourne les statistiques descriptives d'un échantillon ou avec R, utiliser la fonction `quantile` avec l'option `type=1`.

```
quantile(x,type=1)
```

```
##   0%  25%  50%  75% 100%
##   96   98  101  102  104
```

Ou encore pour un petit jeu de données on construit l'échantillon ordonné à la main ici

$$96 \;\; 97 \;\; 98 \;\; 99 \;\; 100 \;\; 101 \;\; 101 \;\; 101 \;\; 102 \;\; 103 \;\; 104 \;\; 104$$

Et l'échantillon étant de taille $n = 12$ on prend la sixième $(n/2)$ valeur pour la médiane et les 3ièmes $(n/4)$ et 9ièmes $(3n/4)$ valeurs pour les premier et troisième quartiles.

question 3

Avec le menu stat. desc. à un échantillon de la calculatrice ou avec R on obtient :

```r
m<-mean(x); var<-sum(x*x)/length(x)-mean(x)**2; et<-sqrt(sum(x*x)/length(x)-mean(x)**2)
c(m,var,et)  # affiche moyenne, variance empirique et écart-type empirique
```

```
## [1] 100.50   6.25   2.50
```

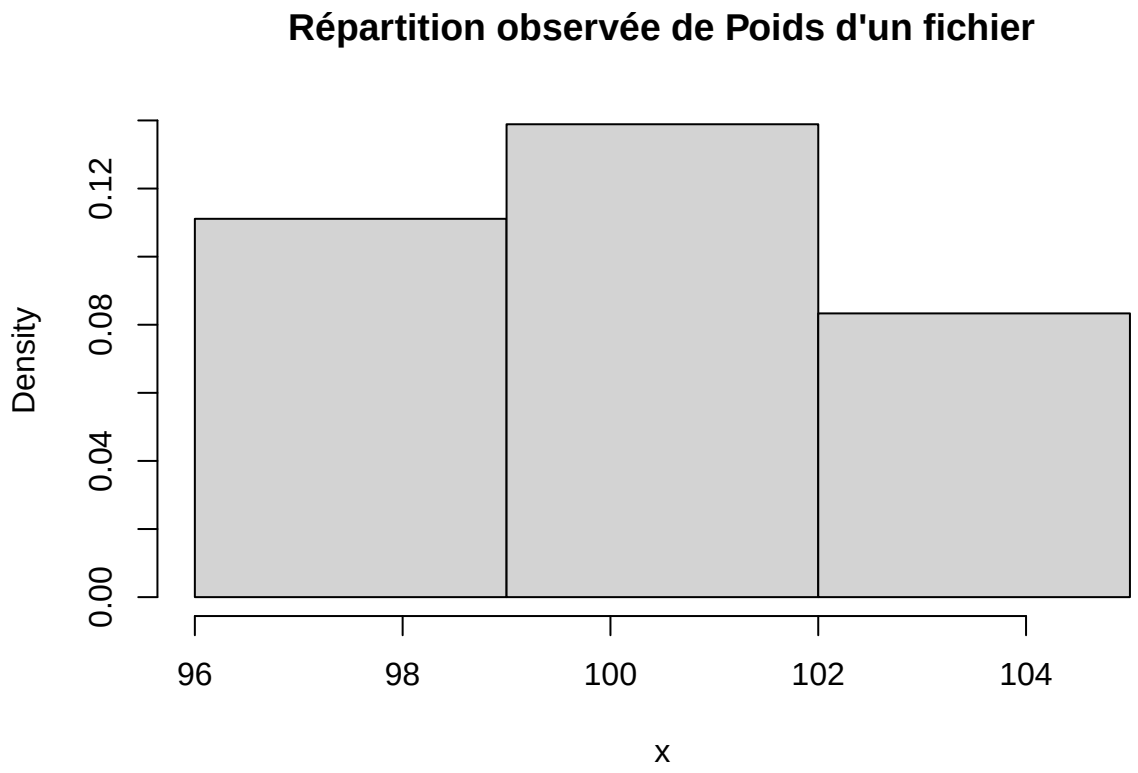
Soit

$$\bar{x} = 100.5 \quad s^2 = 6.25 \text{ et } s = 2.5$$

question 4

On considère les classes $[96, 99]$, $]99, 102]$ et $]102, 105]$ de même longueur. On calcule les effectifs n_k et les fréquences relatives aux largeurs de classe $f_k/(e_k - e_{k-1})$. Avec R on obtient le tracé de l'histogramme avec la fonction `hist()`. Rappelons que chaque rectangle de l'histogramme a pour surface la fréquence qu'il représente. Ainsi la surface du premier rectangle qui vaut $4/12$ est bien $= 0.11 \cdot 3$.

```r
h<-hist(x,breaks=c(96,99,102,105),prob=T,main="Répartition observée de Poids d'un fichier")
```




```r
#trace la répartition observée d'une var continue
h$breaks # les classes
```

```
## [1]  96  99 102 105
```

```r
h$counts #  effectifs de chaque classe
```

```
## [1] 4 5 3
```

```r
h$density #  fréquences relatives aux largeurs de classes
```

```
## [1] 0.11111111 0.13888889 0.08333333
```

```
h$mids # centres des classes
```

```
## [1]  97.5 100.5 103.5
```

question 4

L'échantillon centré réduit est l'échantillon $(y_i)_i$ où $y_i = (x_i - \bar{x})/s$. Puisqu'il est centré, on verifie que sa moyenne empirique vaut 0, et comme il est réduit sa variance empirique vaut 1.

```
y<-(x-mean(x))/sqrt(sum(x*x)/length(x)-mean(x)**2)
y # affiche l'ech centré reduit
```

```
##  [1]  0.2  1.4  0.6 -0.6  0.2 -1.0 -1.4  1.4 -0.2 -1.8  1.0  0.2
```

```
ny<-length(y)
c(mean(y),(ny-1)/ny*var(y)) # affiche moyenne et variance empiriques
```

```
## [1] -9.257505e-18  1.000000e+00
```

On vérifie donc que :

$$\bar{y} = 0 \qquad \text{et} \qquad s_y^2 = 1$$

Exercice 1.3

Le nombre de clics par heure est en moyenne le plus important sur le site B. En revanche il est plus variable sur ce dernier que sur le site A. Autrement dit le nombre de clics à l'heure est moins variable sur A que sur B. En général on compare l'écart-type et la moyenne, car pour un échantillon de moyenne 100 et écart-type 1 la régularité est bien plus basse que pour un échantillon de moyenne 10 et d'écart-type 1 alors que les deux échantillons ont la même variabilité. On comparera donc plutôt les quantités $s/\bar{x}$ pour prendre aussi en compte la valeur de la moyenne. Ainsi ici on a $s_A/\bar{x}_A = 1.8/19.6 \approx 0.09$ et $s_B/\bar{x}_B \approx 0.10$. Comme le ratio "bruit sur signal" est plus faible avec A qu'avec B, on dira que A est plus régulier que B (car avec moins de variabilité de mesure relativement à la moyenne).

Exercice 1.5

On fait les calculs avec R (ou avec la calculette)

```
modalites<-c(3,4,8,11)
effA<-c(6,2,5,9);effB<-c(8,3,11,7);nA<-sum(effA); nB<-sum(effB)
moyA<-sum((modalites*effA))/nA; varA<-sum((modalites**2*effA))/nA-moyA**2;
c(moyA,varA,sqrt(varA)) #affiche moyenne, var et e.t. empiriques de A
```

```
## [1]  7.500000 11.704545  3.421191
```

```
moyB<-sum((modalites*effB))/nB; varB<-sum((modalites**2*effB))/nB-moyB**2;
c(moyB,varB,sqrt(varB)) #affiche moyenne, var et e.t. empiriques de A
```

```
## [1] 6.931034 9.581451 3.095392
```

Donc on obtient

$$\bar{a} = 7.5 \qquad s_A^2 = 11.70 \qquad s_A = 3.4 \qquad \bar{b} = 6.9 \qquad s_B^2 = 9.58 \qquad s_B = 3.1$$

La taille des fichiers est en moyenne la plus importante avec A et elle est la plus variable avec B. Le ratio $s/\bar{x}$ est plus faible pour B (1.3) que pour A (2.2), signe que B est plus régulier que A relativement à sa moyenne.

Exercice 1.6

question 1 :

Notons $x = (x_1, ..., x_{20})$ resp. $y = (y_1, ..., y_{10})$ l'échantillon des notes des filles, resp. des garçons. On donne $\bar{x} = 12$ et $\bar{y} = 8$ et on note $n_x = 20$ et $n_y = 10$ les tailles respectives des échantillons et $n = n_x + n_y$. La moyenne de la classe m s'obtient comme la moyenne des deux moyennes pondérée par les poids n_x/n et n_y/n :

$$m = \frac{1}{n}(\sum x_i + \sum y_i) = \frac{1}{n}(n_x\bar{x} + n_y\bar{y}) = \frac{2}{3}12 + \frac{1}{3}8 = 32/3 \approx 10.7$$

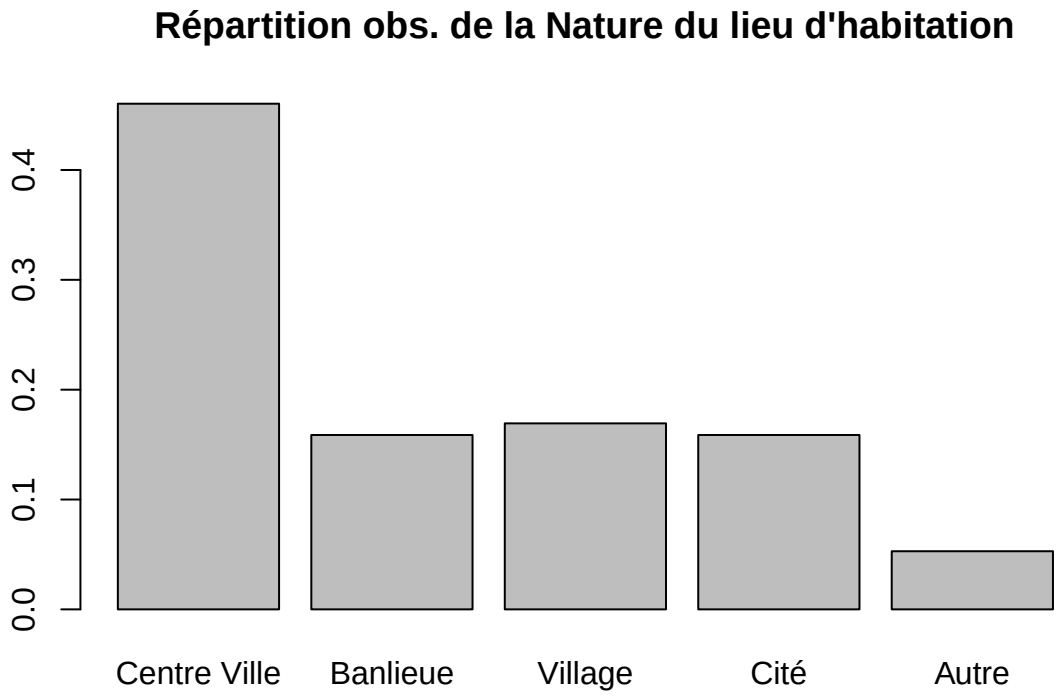
question 2 :

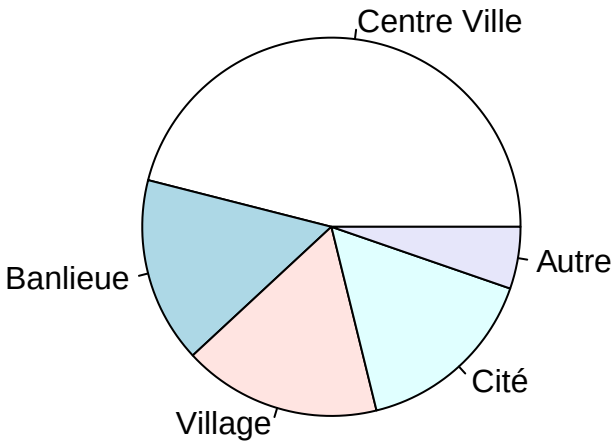
Soit $x = (11, 12, 12, 13, 15, 16, 16, 17, 17, 18, 19, 20, 22, 23)$ l'échantillon qui est déjà ordonné et de taille 14 on prend l'observation qui arrive en position le premier entier supérieur où égal à $n/4$, $n/2$ et $3n/4$ comme quartiles empiriques soit $q_1 = x_{(4)} = 13$, $q_2 = x_{(7)} = 16$ et $q_3 = x_{(11)} = 19$.

question 3 :

Par définition la variable observée est qualitative et a cinq modalités possibles, on peut donc seulement représenter sa répartition observée avec un diagramme en barres ou en secteurs. Aucun autre calcul numérique n'a de sens, donc on ne se pose évidemmet pas la question de calculer médiane, moyenne ou variance. Avec le tronçon suivant d'instructions R on produit le diagramme en barres ou celui en secteurs ("camembert").

```
eff<-c(87,30,32,30,10) # effectifs
noms<-c("Centre Ville", "Banlieue", "Village", "Cité", "Autre") # modalités
freq<-eff/sum(eff) # fréquences
barplot(freq,names.arg=noms, main="Répartition obs. de la Nature du lieu d'habitation"); pie(freq,labels
```





Exercice de la première série de diapositives

On saisit les valeurs des trois échantillons dans `x`, `y` et `z`

```
x<-c(1, 4, 1, 3, 2, 1, 1, 4, 3, 1, 4, 1, 3, 3, 2)
y<-c(3, 3, 4, 2, 1, 3, 3, 2, 2, 3, 1, 1, 3, 3, 4)
z<-c(2, 3, 3, 1, 4, 4, 3, 1, 4, 4, 2, 4, 4, 2, 4, 4, 2, 3, 3, 3)
```

On calcule les effectifs, les fréquences et les fréquences cummulées avec la fonction `table`. Les trois variables sont clairement discrètes.

Pour `x`:

```
table(x); #modalités et effectifs pour x
```

```
## x
## 1 2 3 4
## 6 2 4 3
```

```
freqx<-table(x)/sum(table(x)); freqx #les fréquences pour x
```

```
## x
##         1         2         3         4
## 0.4000000 0.1333333 0.2666667 0.2000000
```

```
cumfreqx<-cumsum(freqx); cumfreqx #les F_k pour x
```

```
##         1         2         3         4
## 0.4000000 0.5333333 0.8000000 1.0000000
```

```r
mean(x); mean(x*x)-mean(x)^2; sqrt(mean(x*x)-mean(x)^2) # moy emp, var emp et etemp pour x
```

```
## [1] 2.266667
```

```
## [1] 1.395556
```

```
## [1] 1.181336
```

Dans le tableau décrivant x de la diapositive n°44 on complète les valeurs successives manquantes pour la ligne des fréquences avec : 13.33, 20 et 100 (total). Puis pour celles des fréquences cumulées (exprimées en pourcentages) et de gauche à droite : 40, 53.33 et 100 (modalité 4) et la cellule total sur la ligne des fréquences cumulées n'a pas de sens et doit être grisée. Par contre il est important de s'assurer que la fréquence cumulée de la plus grande modalité vaut bien 100%.

Pour obtenir les résumés numériques avec R on utilise `summary()` sur chacun des échantillons.

```r
summary(x); summary(y); summary(z)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   2.267   3.000   4.000
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.533   3.000   4.000
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1       2       3       3       4       4
```

Cela fournit toutes les valeurs manquantes dans le tableau des résumés numériques de la diapositive n°44, à l'exception des trois variances empiriques (et leurs racines) qu'on peut calculer (et afficher) à part avec

```r
c(mean(x*x)-mean(x)^2,mean(y*y)-mean(y)^2,mean(z*z)-mean(z)^2) #calc et aff des var emps
```

```
## [1] 1.3955556 0.9155556 1.0000000
```

```r
sqrt(c(mean(x*x)-mean(x)^2,mean(y*y)-mean(y)^2,mean(z*z)-mean(z)^2)) #et emps.
```

```
## [1] 1.1813363 0.9568467 1.0000000
```