TD1

ING1 EC551 Statistiques descriptives

Galharret Jean-Michel
département MSC
https://galharret.github.io/WEBSITE/cours_ONIRIS.html

Exercice 1 :

Dans le cadre d'un programme de gestion de la production, des données relatives au temps nécessaire pour accomplir une tâche T ont été relevées sur 50 jours (données exprimées en heures*homme) :

```
x<-c(128,119,95,97,124,128,142,98,108,120,
        113,109,124,132,97,138,133,136,120,112,
        146,128,103,135,114,109,100,111,131,113,
        124,131,133,131,88,118,116,98,112,138,
        100,112,111,150,117,122,97,116,92,112)
```

1. Identifiez la variable et sa nature.

La variable est quantitative continue il s'agit du temps d'accomplissement de la tâche.

2. Calculer la moyenne et l'écart type de la série de données. On les note $\bar{x}, \sigma(x)$.

```
mean(x)
```

```
[1] 117.62
```

```
sd(x)
```

```
[1] 15.02092
```

3. On décide de regrouper les données en classe de longueur 10 (fonction ***cut*** avec comme argument ***breaks*** dans R). En déduire la table des effectifs obtenus. Ajouter à cette table la table des fréquences (fonction proportions dans R).

```
min(x)
```

```
[1] 88
```

```
max(x)
```

```
[1] 150
```

```
z<-cut(x,breaks=seq(80,150,10))
effectifs<-table(z)
frequences<-proportions(effectifs)
```

4. Recalculer la moyenne et l'écart type pour les données réparties en classe, on a les formules suivantes :

- Moyenne $\bar{c} = \sum_{i=1}^{I} f_i c_i$ où I est le nombre de classe utilisées et c_i le centre de chacune d'elles.
- Ecart type $\sigma(c) = \sqrt{\sum_{i=1}^{I} f_i (c_i - \bar{c})^2}$ où I est le nombre de classe utilisées et c_i le centre de chacune d'elles.

On calcule le centre des classe :

```
centres<-seq(85,145,10)
```

On en déduit que la moyenne est égale à

```
moy<-sum(frequences*centres)
moy
```

```
[1] 117.2
```

et l'écart type vaut

```
sqrt(sum(frequences*(centres-moy)^2))
```

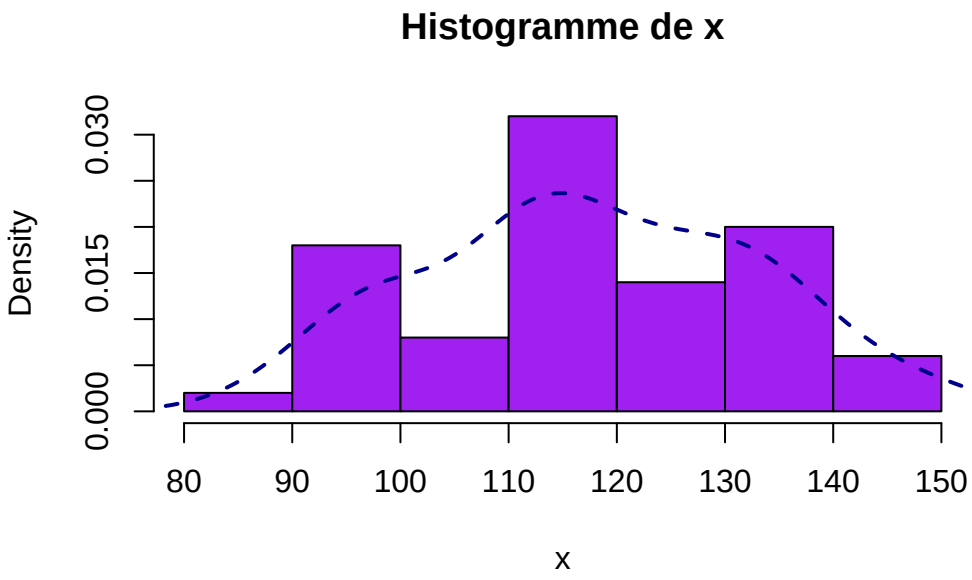
```
[1] 15.52933
```

5. Comparer les valeurs obtenues pour $\bar{x}, \bar{c}$ et $\sigma(x), \sigma(c)$. Quel paramètre du regroupement en classe faudrait-il ajuster pour que $\bar{c}$ soit une meilleure approximation de $\bar{x}$?

Réponse : Le fait d'avoir regroupé les valeurs en classes entraine une perte d'information mais permet néanmoins d'obtenir des estimations de la moyenne des observations et de l'écart type proches des vrais valeurs.

6. Tracer l'histogramme de la série de données, on prendra le paramètre par défaut puis on prendra ***breaks=k*** où $k = 2.5 \times N^{.25}$ (formule de Yule). Ajouter sur chacun des histogrammes la fonction densité avec ***lines*** et ***density***. Que peut-on dire de ces deux représentations graphiques.

```
hist(x,probability = T,col="purple",main="Histogramme de x")
lines(density(x),col="darkblue",lty=2,lwd=2)
```



Exercice 2 :

On considère le fichier contenant les huiles d'olive. Télécharger ce fichier sur CONNECT. Ouvrez le fichier dans R.

1. On veut construire la table de fréquence de la provenance des huiles d'olive en fonction de la région. On utilisera la fonction ***proportions***.

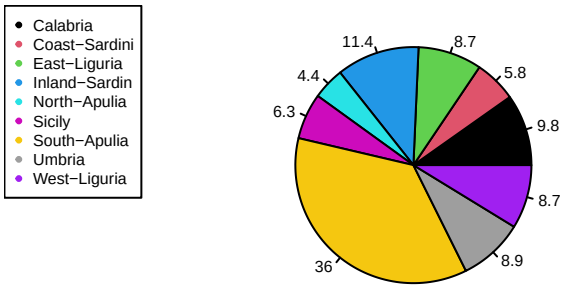
```
library(readr)
h_olive <- read_delim("~/Dropbox/ONIRIS/cours/EC551_Stat_Desc/TD/h_olive.csv",
    ";", escape_double = FALSE, trim_ws = TRUE)
```

```
Rows: 572 Columns: 10
-- Column specification --------------------------------------------------------
Delimiter: ";"
chr (1): Region
dbl (9): ID, Palmitic, Palmitoleic, Stearic, Oleic, Linoleic, Eicosanoic, Li...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
effectifs<-table(h_olive$Region)

frequences<-proportions(effectifs)
```

2. Reproduire le diagramme circulaire du cours (diapo 19).

```
h_olive$Region<-as.factor(h_olive$Region)
pie(frequences,labels=round(100*frequences,1),cex=.5,col=c(1:8,"purple"))
legend("topleft",levels(h_olive$Region),col=c(1:8,"purple"),pch=rep(20,9),cex=.5)
```



3. En vous inspirant de la diapositive 22 créer une variable ordinale à partir de la variable Oleic ayant pour modalité :

- TF (très faible) pour $Oleic < 6500$,
- F pour $6500 \leq Oleic < 7400$,
- M pour $7400 \leq Oleic < 8000$,
- Fo pour $Oleic \geq 8000$.

Représenter cette variable par le graphique adéquat. On calculera la table des fréquences.

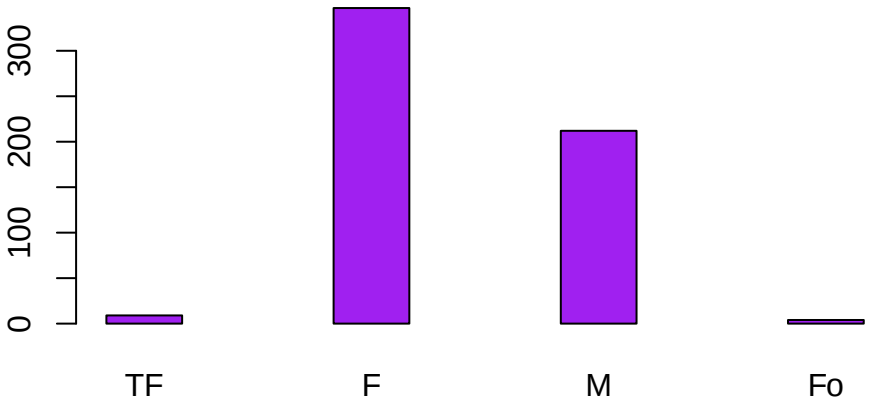
```
min(h_olive$Oleic)
```

```
[1] 6300
```

```
max(h_olive$Oleic)
```

```
[1] 8410
```

```
x<-cut(h_olive$Oleic,breaks=c(6200,6500,7400,8000,8500))
z<-factor(x,labels=c("TF","F","M","Fo"))
barplot(table(z),col="purple",space=2)
```



Attention : comme il s'agit d'une variable ordinale on n'utilise pas un diagramme circulaire.

4. Créer une fonction notée SD permettant de calculer l'écart type d'une série d'observations x (vecteur). Quelle est la différence avec la fonction *sd* inclue dans R ?

```
VAR<-function(x){
  return( mean((x-mean(x))^2) )
}
```

La fonction ***var*** de R calcule une estimation de l'écart type de la population (divise par $N-1$) alors que ***VAR*** calcule l'écart type des observations.

a. Calculer les écarts types de la série de données correspondantes à la variable Oleic et celle de la variable Palmitic.

b. Créer une fonction Standardise qui étant donné un vecteur x et un indice i renvoie la valeur $z_i = \frac{x_i - \bar{x}}{\sigma(x)}$. Utiliser cette fonction pour l'huile à l'indice 132. Est-elle plus typique sur l'acide palmitique ou bien sur l'acide oléique ?

```
Standardise<-function(x,i){
  return(z<-(x[i]-mean(x))/sqrt(VAR(x)))
}
```

5. Créer trois fonctions permettant de retourner les valeurs atypiques et leur indice (diapo 65).

```
out1<-function(x){
  q1<-quantile(x,probs=.25)
  q3<-quantile(x,probs=.75)
  iqr<-q3-q1
  z<- (x<q1-1.5*iqr) | (x>q3+1.5*iqr)
  return(list(indices=which(z==TRUE),valeurs=x[z]))
}

out2<-function(x){
  z<- abs(x-mean(x))>3*sd(x)
  return(list(indices=which(z==TRUE),valeurs=x[z]))
}

out3<-function(x){
  z<- abs(x-median(x))>4.55*mad(x,constant=1)
  return(list(indices=which(z==TRUE),valeurs=x[z]))
}
```

6. Tester ces trois fonctions sur la série de données Oleic. Que peut on en déduire ?

```
out1(h_olive$Oleic)
```

```
$indices
integer(0)

$valeurs
numeric(0)
```

```
out2(h_olive$Oleic)
```

```
$indices
integer(0)

$valeurs
numeric(0)
```

```
out3(h_olive$Oleic)
```

```
$indices
integer(0)

$valeurs
numeric(0)
```