devoir_MameThierno_exo_rg

Mame Thierno Ndiaye

2023-04-26

Statistique univarié

fonction d.var.quant

Nous allons créer une fonction d.var.quant qui décrie une variable quantitave.La fonction affichera les tendances centrales, les graphiques (hist,boxplot, etc. . .), intervalle de confiance.

```
 d.var.quant<- function(variable){
    Des=summary(variable)
  Graphe=hist(variable, main = "Histogramme des données", xlab = "Variable",
ylab = "Fréquence")
 Plot=boxplot(variable, main = "Boite à moustache", ylab = "Valeurs")
 library(questionr)
 Plot
 Graphe
 print(Des)
 t.test(variable)
 }
```

fonction d.var.quali

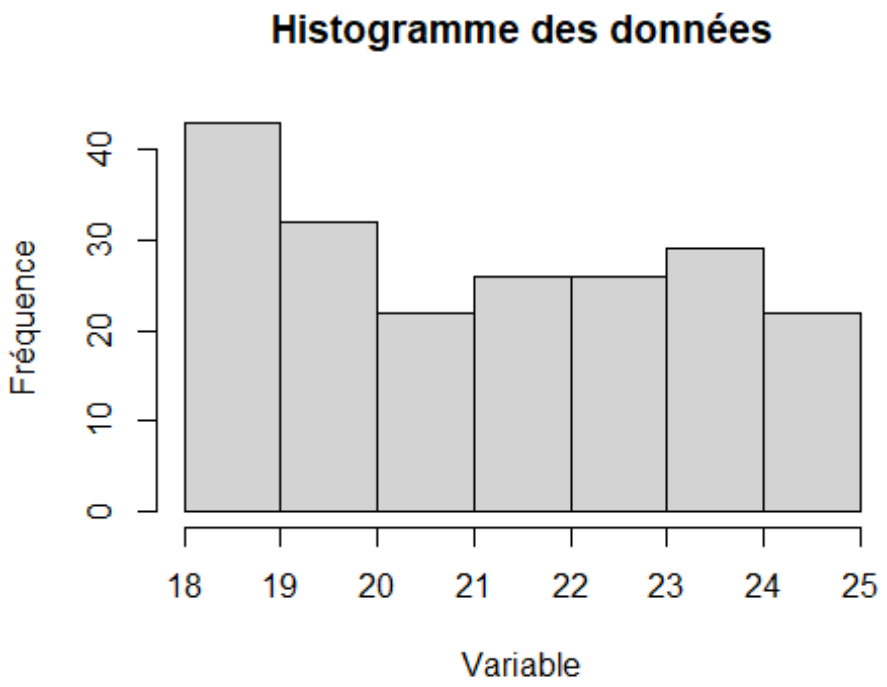
Nous allons créer une fonction d.var.quali qui décrie une variable qualitative.La fonction affichera la frequence, les graphiques (Graphique en secteurs,Graphique en barres)

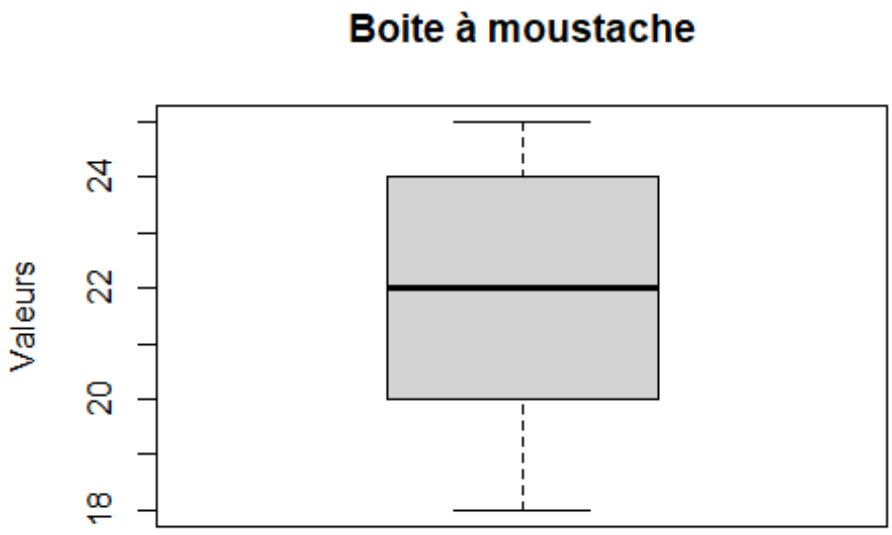
```
 d.var.quali<- function(variable){
  f=data.frame(table(variable))
barplot(f$Freq, names.arg = f$variable, main = "Graphique en barres des
données", xlab = "Données", ylab="Frequence")
 pie(f$Freq, labels =f$variable, main = "Graphique en secteurs des données")
  table(variable)
 }
```

Application

Nous allons maintenant appliquer les deux fonctions créées précedement sur les variables de notre base de données.

```
#Pour les variables quantitative:
d.var.quant(df.MameThiernoNdaiye$age)
```

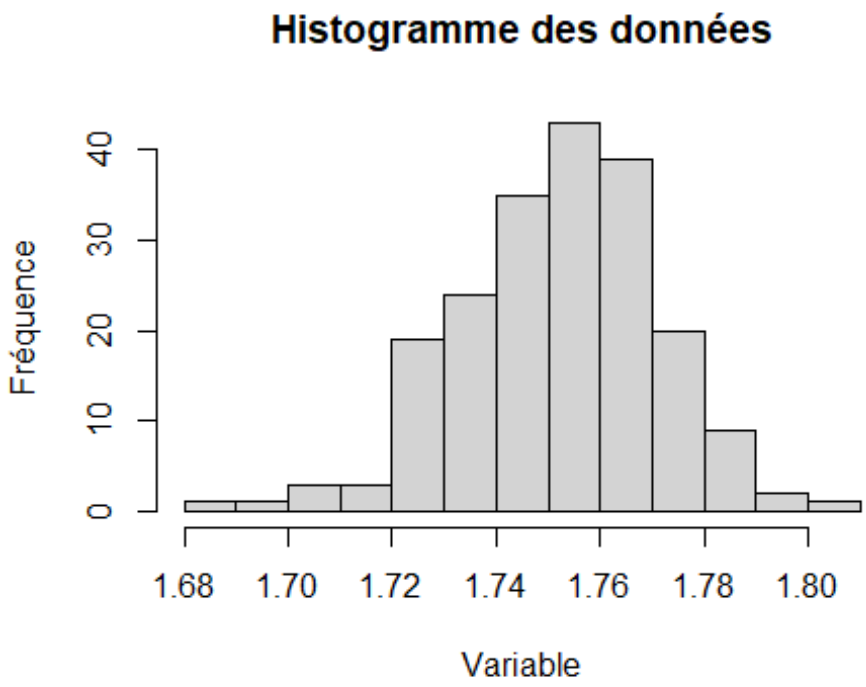


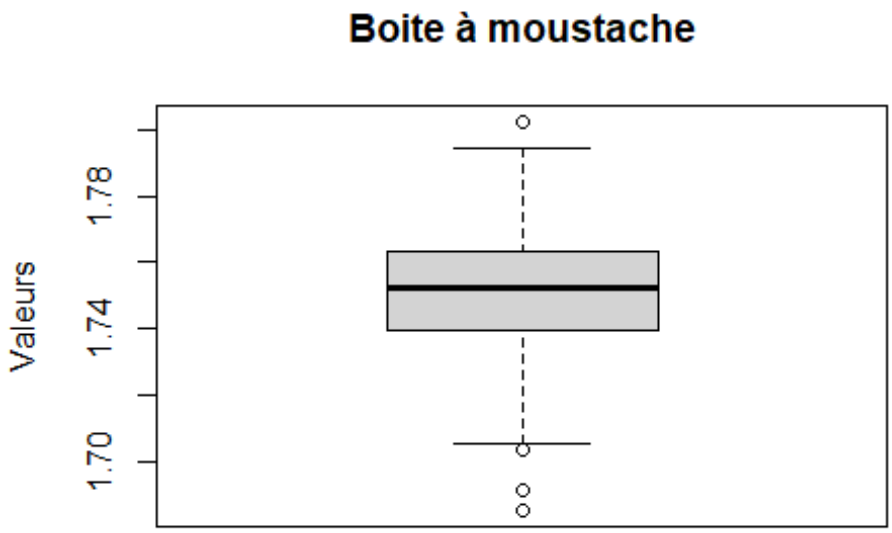



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   20.00   22.00   21.59   24.00   25.00
```

```
##
##  One Sample t-test
##
## data:  variable
## t = 138.84, df = 199, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  21.28336 21.89664
## sample estimates:
## mean of x
##     21.59
```

```
d.var.quant(as.numeric(df.MameThiernoNdaiye$Taille))
```

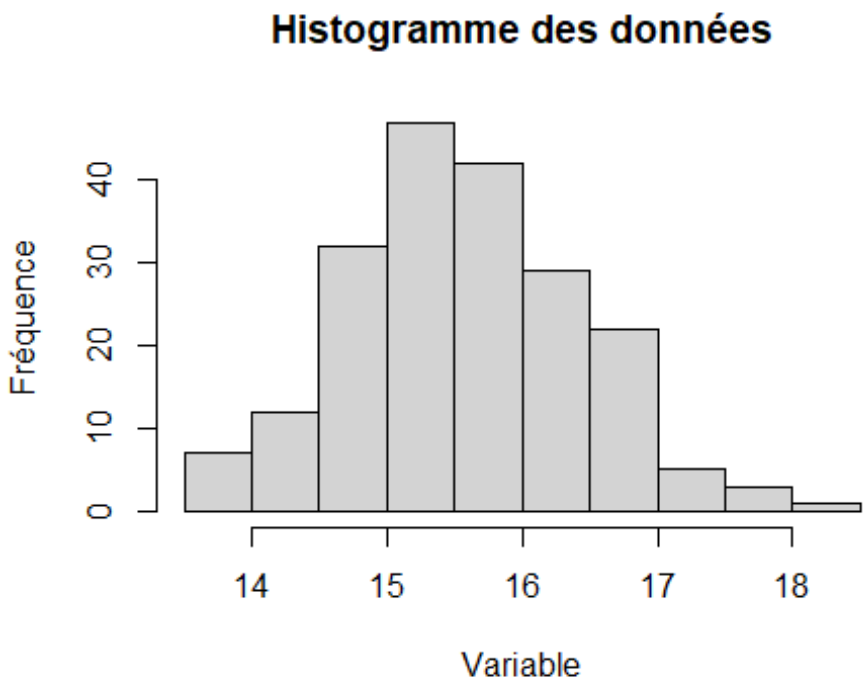


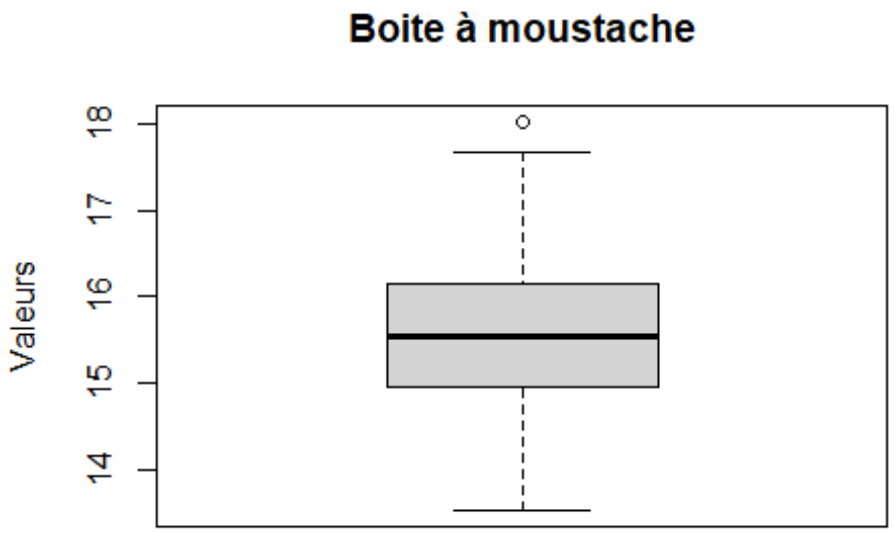



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.685   1.740   1.753   1.752   1.764   1.802
```

```
##
##  One Sample t-test
##
## data:  variable
## t = 1289.5, df = 199, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.748907 1.754265
## sample estimates:
## mean of x
##  1.751586
```

```
d.var.quant(as.numeric(df.MameThiernoNdaiye$Moy_seme1))
```

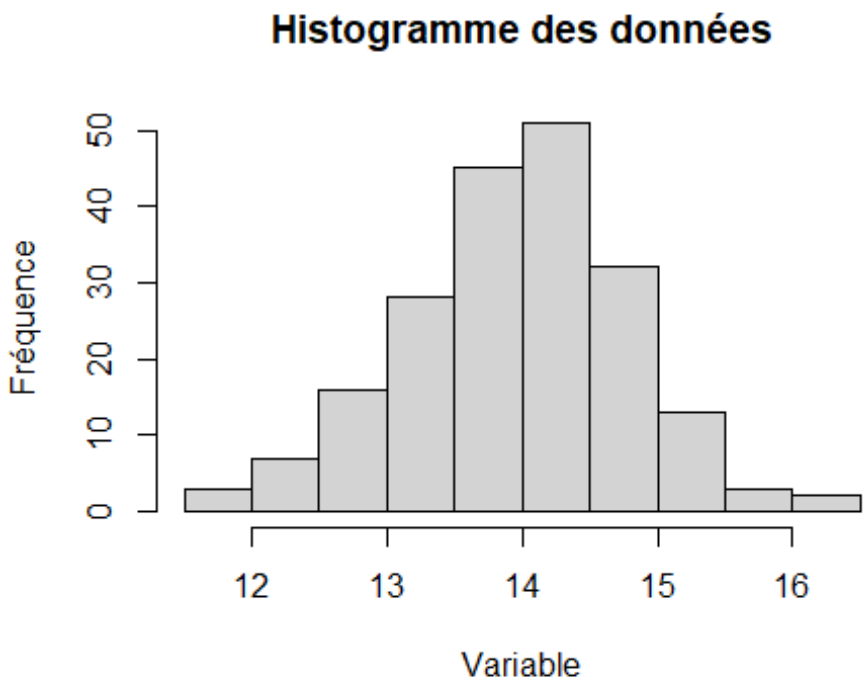


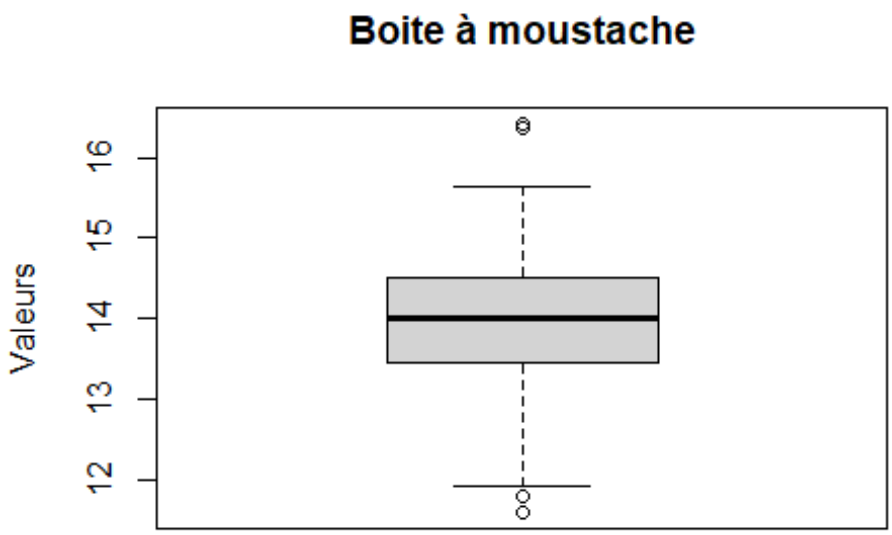



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.54   14.98   15.53   15.57   16.15   18.02
```

```
##
##  One Sample t-test
##
## data:  variable
## t = 258.97, df = 199, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  15.44879 15.68587
## sample estimates:
## mean of x
##  15.56733
```

```
d.var.quant(as.numeric(df.MameThiernoNdaiye$Moy_final))
```

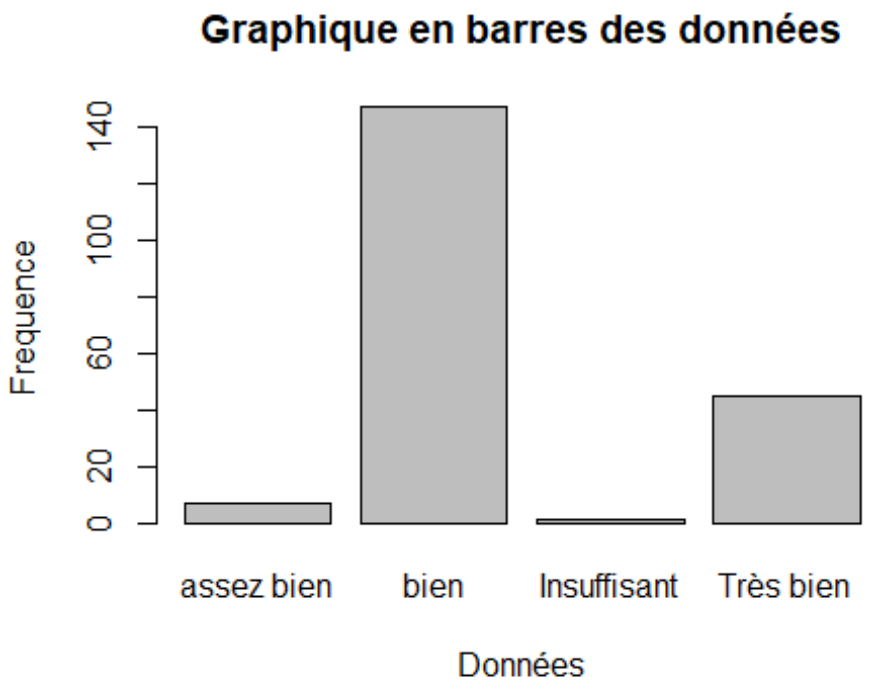


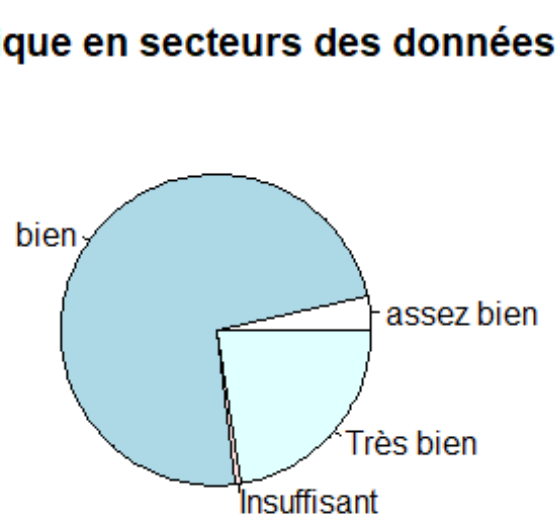



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.61   13.46   14.01   13.94   14.50   16.42
```

```
##
##  One Sample t-test
##
## data:  variable
## t = 234.72, df = 199, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  13.8273 14.0616
## sample estimates:
## mean of x
##  13.94445
```

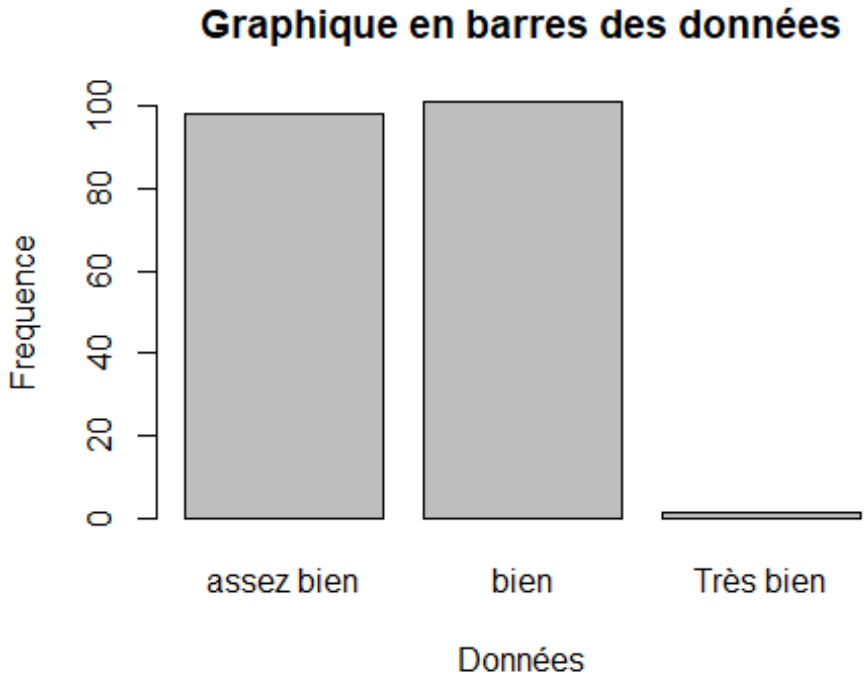
```
#Pour les variables qualitative:
d.var.quali(df.MameThiernoNdaiye$mention_Seme1)
```

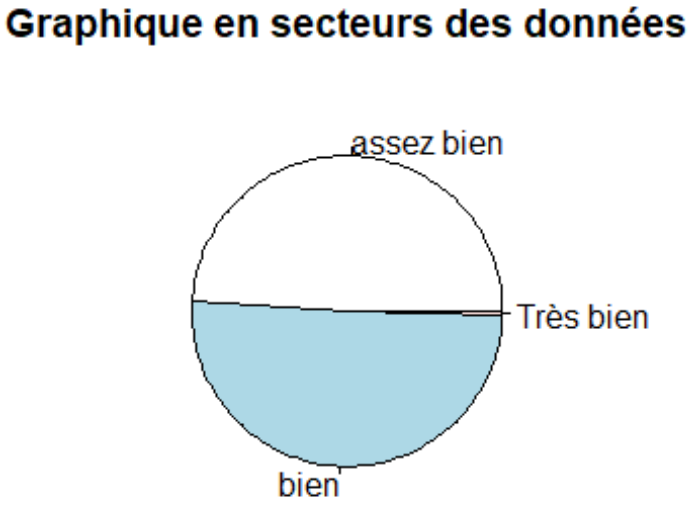




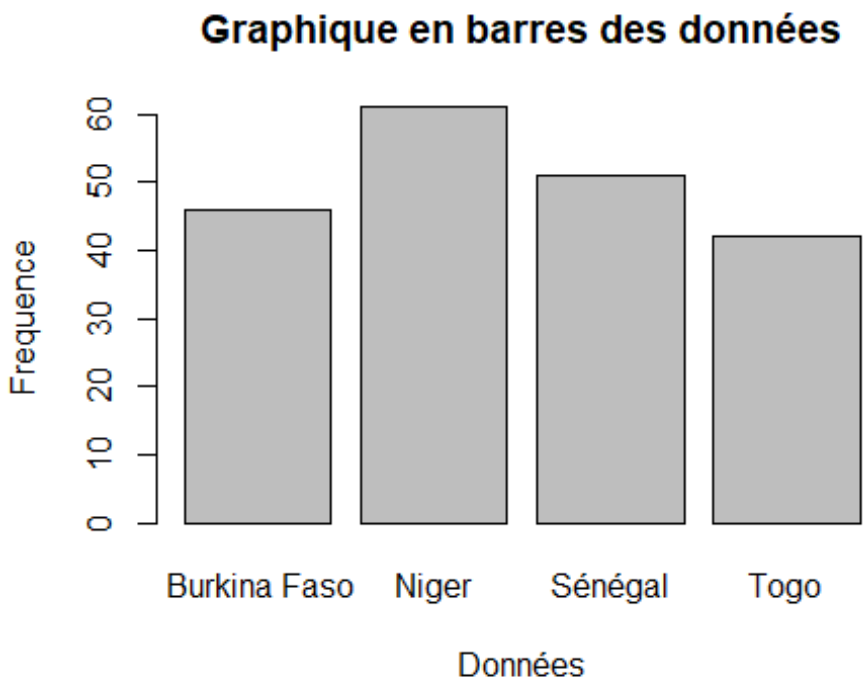

```
## variable
##  assez bien        bien Insuffisant   Très bien
##           7         147           1          45
```

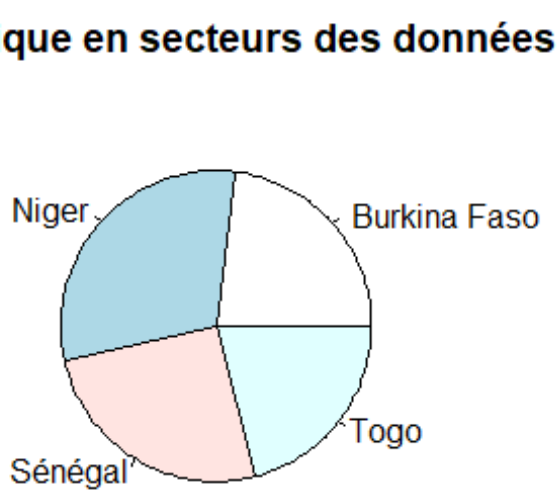
```
d.var.quali(df.MameThiernoNdaiye$mention_final)
```





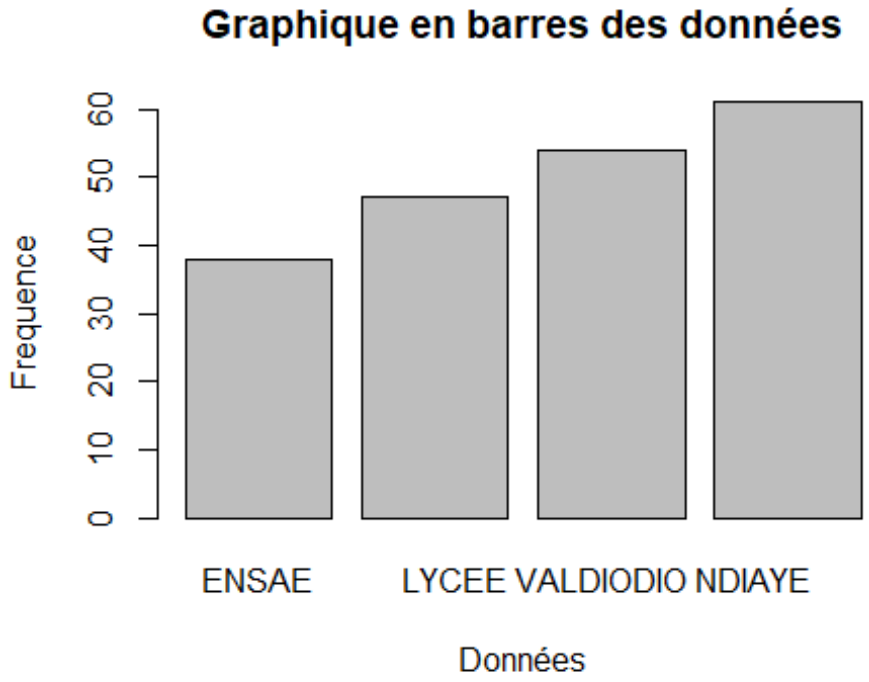
```
## variable
## assez bien       bien  Très bien
##         98        101          1

d.var.quali(df.MameThiernoNdaiye$nationalité)
```

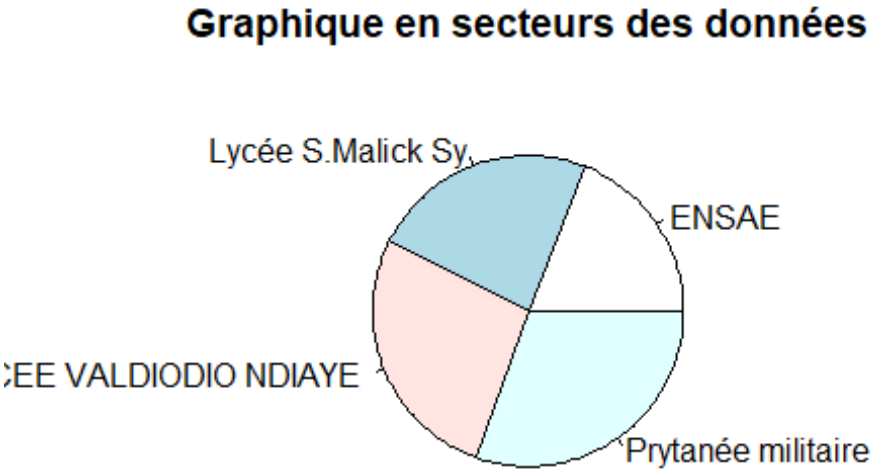





```
## variable
##  Burkina Faso         Niger       Sénégal          Togo
##            46            61            51            42
```

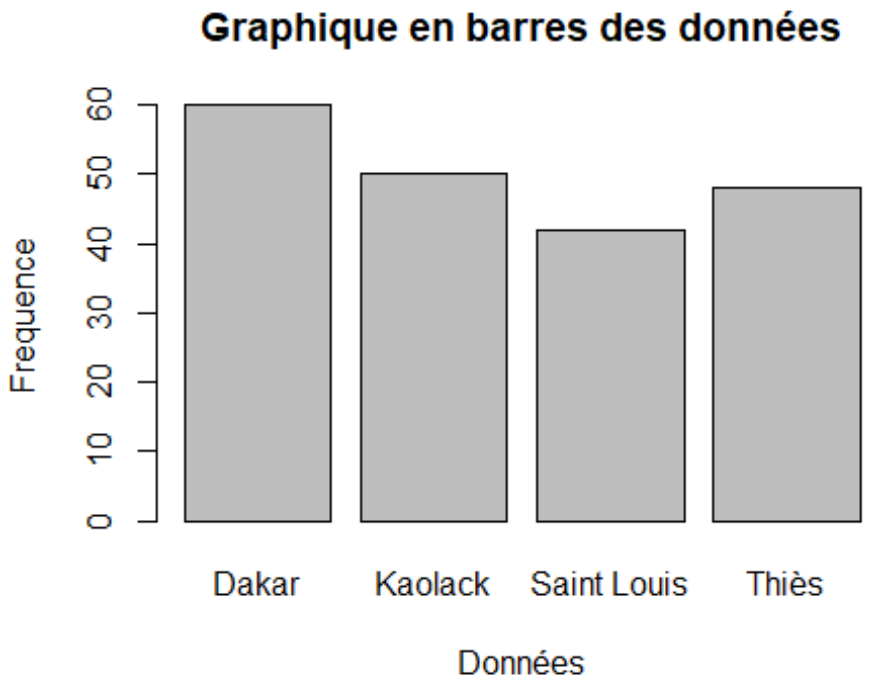
```
d.var.quali(df.MameThiernoNdaiye$centre_examen)
```

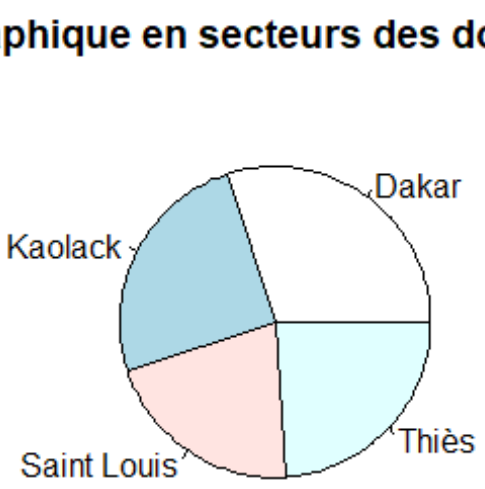





```
## variable
##                   ENSAE       Lycée S.Malick Sy LYCEE VALDIODIO NDIAYE
```

```
##                       38                       47                       54
##      Prytanée militaire
##                       61

d.var.quali(df.MameThiernoNdaiye$Région_Examen)
```





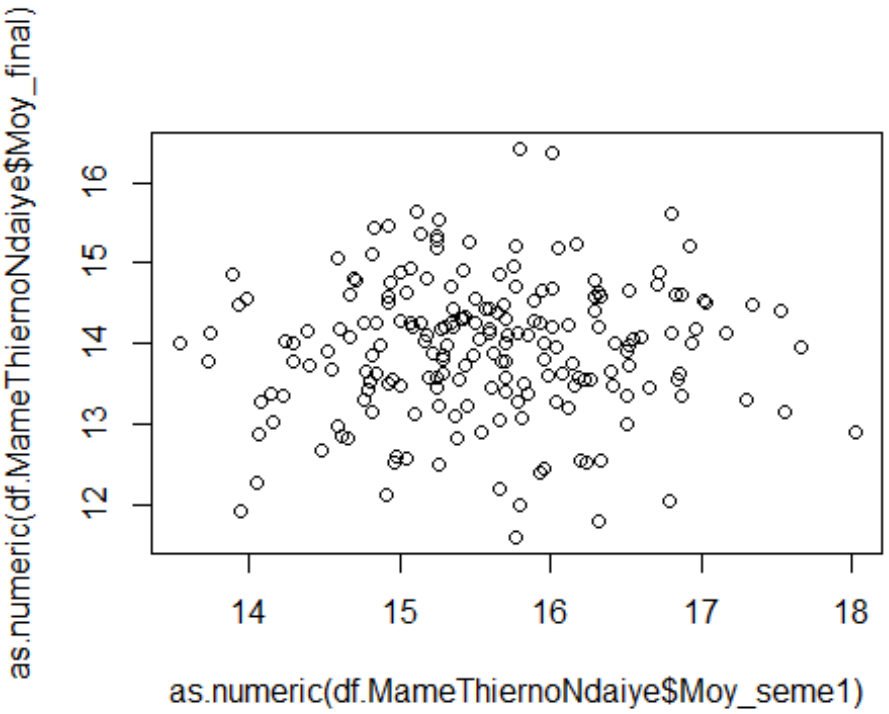

```
## variable
##       Dakar     Kaolack Saint Louis       Thiès
##          60          50          42          48
```

Statistique bivariée

Liaison entre deux variables quantitaives:

Nous allons etudier le lien entre deux variables quantitatives. On prendra la **Moy_seme1** et la **Moy_final**. on va d'abord reprensenter le nuage de points entre les deux variables puis calculer la Corrélation linéaire entre les deux variables.

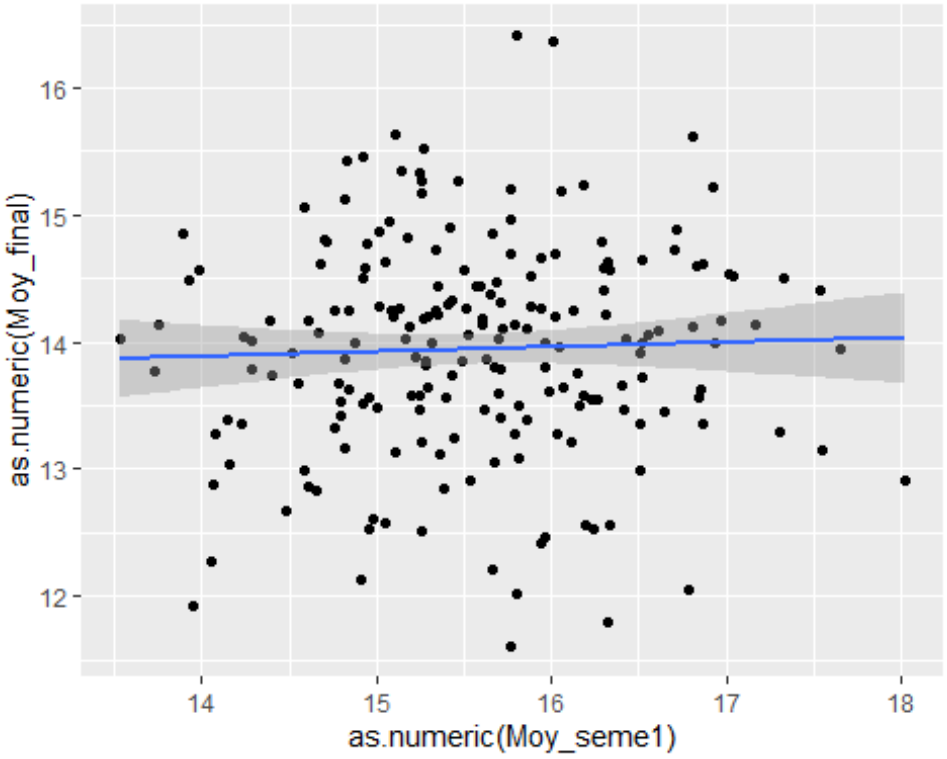
```
plot(as.numeric(df.MameThiernoNdaiye$Moy_seme1) ,
as.numeric(df.MameThiernoNdaiye$Moy_final))
```




```
cor(as.numeric(df.MameThiernoNdaiye$Moy_seme1) ,
as.numeric(df.MameThiernoNdaiye$Moy_final))

## [1] 0.03720194
```

Les resultats obtenus montrent qu'il y'a pas une forte relation lineaire entre les deux variables. on va maintenant représenter la droite de régression sur notre nuage de points.

```
library(ggplot2)
ggplot(df.MameThiernoNdaiye, aes(x=as.numeric(Moy_seme1),
y=as.numeric(Moy_final))) +
  geom_point()+
  geom_smooth(method=lm)

## `geom_smooth()` using formula = 'y ~ x'
```



Liaison entre deux variables qualitative:

Nous allons etudier le lien entre deux variables quantitatives. On prendra la **nationalité** et la **mention_final**. on va d'abord faire le tableau croisé des deux variables.

```
library(questionr)
#Tableau en pourcentages ligne ou colonne.
tab=table(df.MameThiernoNdaiye$nationalité,
df.MameThiernoNdaiye$mention_final)
lprop(tab)
```

```
##
##                 assez bien bien  Très bien Total
##    Burkina Faso  39.1       60.9   0.0     100.0
##   Niger          50.8       49.2   0.0     100.0
##   Sénégal        52.9       47.1   0.0     100.0
##   Togo           52.4       45.2   2.4     100.0
##   Ensemble       49.0       50.5   0.5     100.0
```

```
cprop(tab)
```

```
##
##                 assez bien bien  Très bien Ensemble
##    Burkina Faso  18.4       27.7   0.0      23.0
##   Niger          31.6       29.7   0.0      30.5
##   Sénégal        27.6       23.8   0.0      25.5
```

```
##   Togo            22.4       18.8 100.0      21.0
##   Total          100.0      100.0 100.0     100.0
```

On va maintenant proceder au Test du χ^2

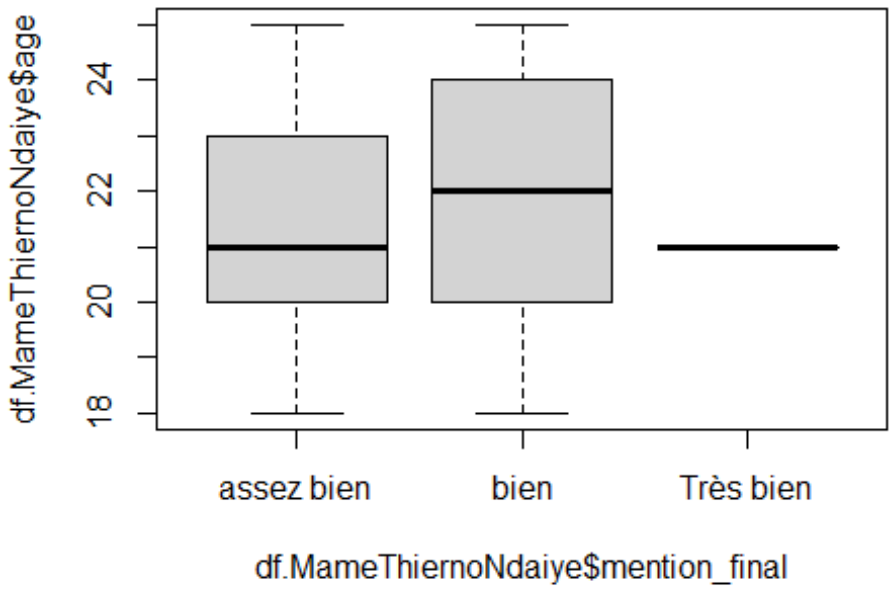
```
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 6.3276, df = 6, p-value = 0.3875
```

```
#X-squared: c'est la valeur de la statistique du χ² pour notre tableau,
c'est-à-dire une "distance" entre notre tableau observé et celui attendu si
les deux variables étaient indépendantes.
#df: le nombre de degrés de libertés du test, qui dépend des dimensions du
tableau.
#p-value:C'est la probabilité d'obtenir une valeur de la statistique du χ² au
moins aussi extrême sous l'hypothèse d'indépendance..
```

Les resultats du test montre un p-value = 0.3875, donc il n'y pas de relation entre les deux variables. # Liaison entre Une variable qualitative et une var quantitative Nous allons etudier le lien entre Une variable qualitative et une variable quantitative. On prendra la **l'âge** et la **mention_final**. on va d'abord faire une représentation graphique des boites à moustaches des deux variables.

```
boxplot(df.MameThiernoNdaiye$age ~ df.MameThiernoNdaiye$mention_final)
```



Dans le graphique généré, on voit que ceux qui ont une mention assez bien semble être au même age que ceux qui ont la mention bien. Pour verifier notre intuition, On va maintenant calculer la moyenne d'âge pour les differentes mentions.

```
library(dplyr)

##
## Attachement du package : 'dplyr'

## Les objets suivants sont masqués depuis 'package:stats':
##
##     filter, lag

## Les objets suivants sont masqués depuis 'package:base':
##
##     intersect, setdiff, setequal, union

tapply(df.MameThiernoNdaiye$age, df.MameThiernoNdaiye$mention_final, mean)

## assez bien       bien  Très bien
##   21.52041   21.66337   21.00000
```

La mention T_bien n'a été obtenue que par une seule personne et donc ne peut pas nous donner beaucoup d'information. Si on prend les mention bien et assez bien, les resultats montre que la moyenne d'âge chez les mentions bien est sensiblement la m$emee chez les assez bien.

on va maintenant proceder aux test pour confirmer ou pas notre intuition.

```
d=filter(df.MameThiernoNdaiye, mention_final != "Très bien" )
t.test (d$age ~ d$mention_final)

##
##  Welch Two Sample t-test
##
## data:  d$age by d$mention_final
## t = -0.45698, df = 196.65, p-value = 0.6482
## alternative hypothesis: true difference in means between group assez bien
and group bien is not equal to 0
## 95 percent confidence interval:
##  -0.7598902  0.4739738
## sample estimates:
## mean in group assez bien       mean in group bien
##                 21.52041                 21.66337
```

Les resultats donnent un p_value=0.6482.Ceci montre qu'il y'a pas de relation etre les deux variables.