Basic Concept of Statistics

Paolo Girardi and Livio Finos

8/10/2020

Contents

Descriptive Statistics

Univariate Statistical Analysis with R

We import the dataset test.csv

```
test<-read.csv("test.csv",sep=";",header=T,dec=",")
head(test) #the first 6 rows
```

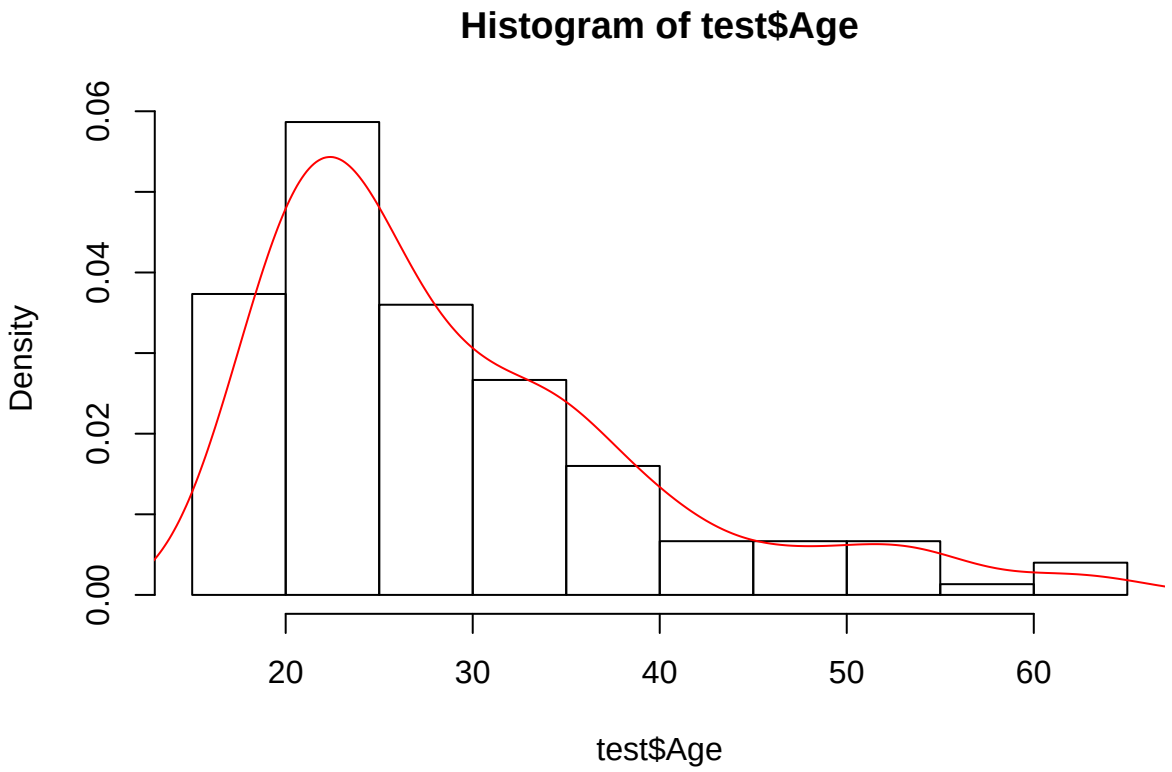
```
##   ID Age  BMI Gender     Education ACT SATV SATQ Stress Social
## 1  1  19 24.3      F     secondary  24  500  500      2      3
## 2  2  23 24.6      F     secondary  35  600  500      1      6
## 3  3  20 28.1      F     secondary  21  480  470      6      2
## 4  4  27 24.5      M        degree  26  550  520      1      3
## 5  5  33 24.1      M upper primary  31  600  550      5      2
## 6  6  26 23.1      M   post-degree  28  640  640      6      1
```

```
str(test)
```

```
## 'data.frame':    150 obs. of  10 variables:
##  $ ID       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Age      : int  19 23 20 27 33 26 30 19 23 40 ...
##  $ BMI      : num  24.3 24.6 28.1 24.5 24.1 23.1 23.2 21.9 27.3 24.1 ...
##  $ Gender   : Factor w/ 2 levels "F","M": 1 1 1 2 2 2 1 2 1 1 ...
##  $ Education: Factor w/ 6 levels "degree","lower primary",..: 5 5 5 1 6 3 3 5 1 3 ...
##  $ ACT      : int  24 35 21 26 31 28 36 22 22 35 ...
##  $ SATV     : int  500 600 480 550 600 640 610 520 400 730 ...
##  $ SATQ     : int  500 500 470 520 550 640 500 560 600 800 ...
##  $ Stress   : int  2 1 6 1 5 6 5 4 4 4 ...
##  $ Social   : int  3 6 2 3 2 1 5 2 6 5 ...
```

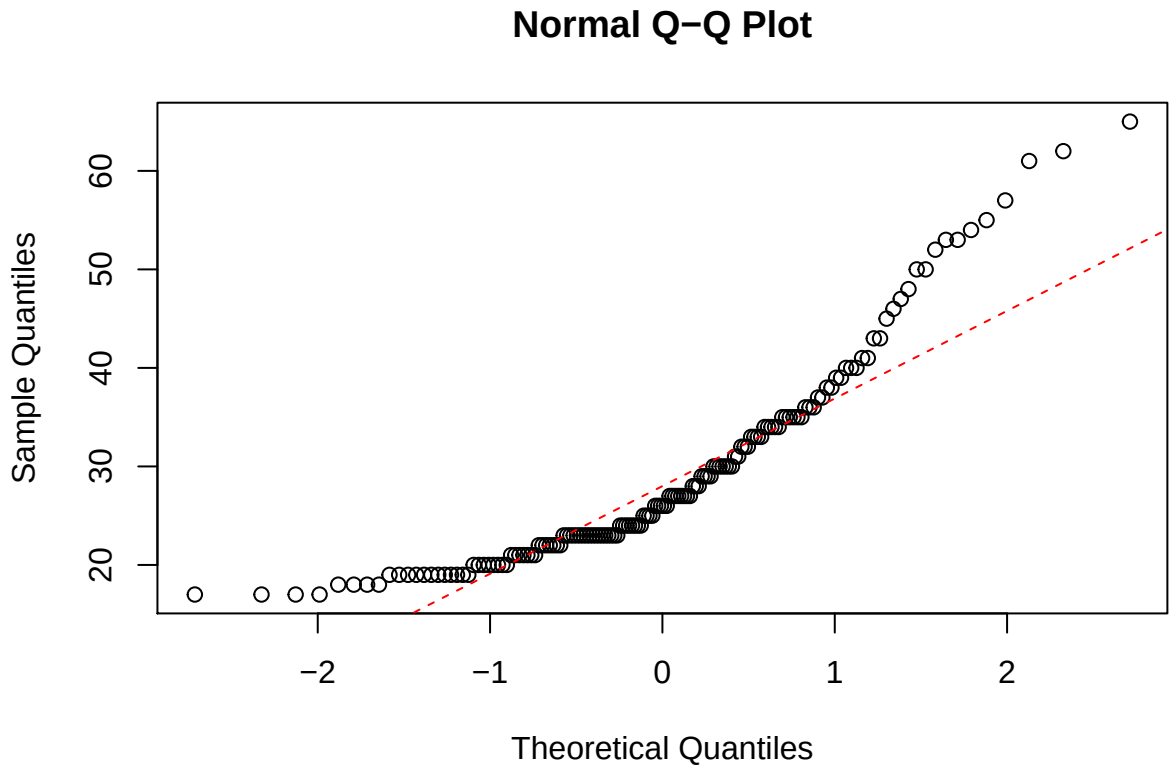
Analysis of the Age variable

```
# A histogram with the density plot
hist(test$Age,prob=T)
lines(density(test$Age),col=2)
```



The distribution is skewed. A qqplot can be used to visualise the distribution

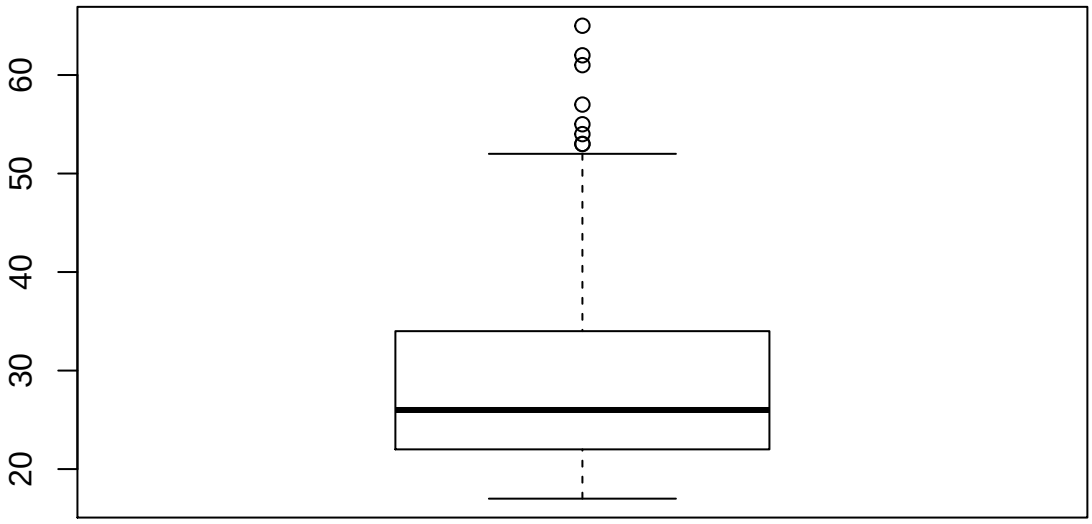
```
qqnorm(test$Age)
qqline(test$Age,col=2,lty=2)
```



The graph reports the comparison between the theoretical quantile of a Normal distribution and quantiles of the variabile Age. If the points follow the red line, a normal distribution can be assumed.

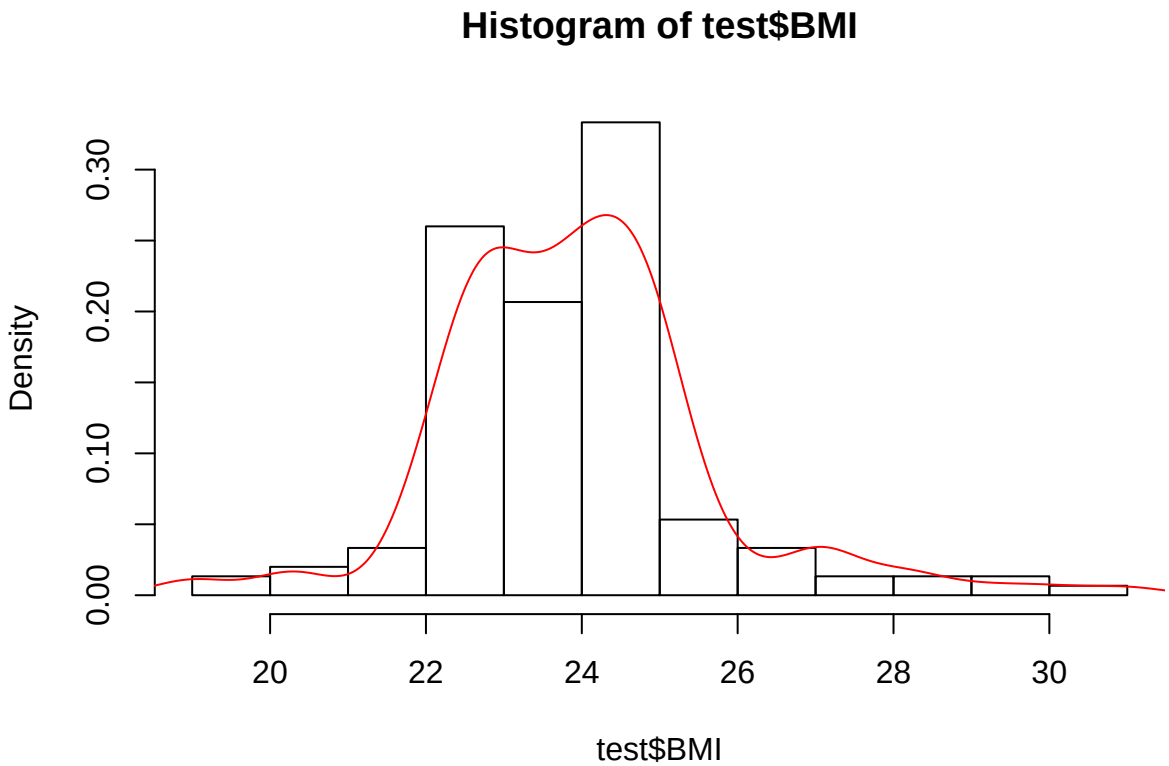
The function *boxplot()*

```
boxplot(test$Age)
```



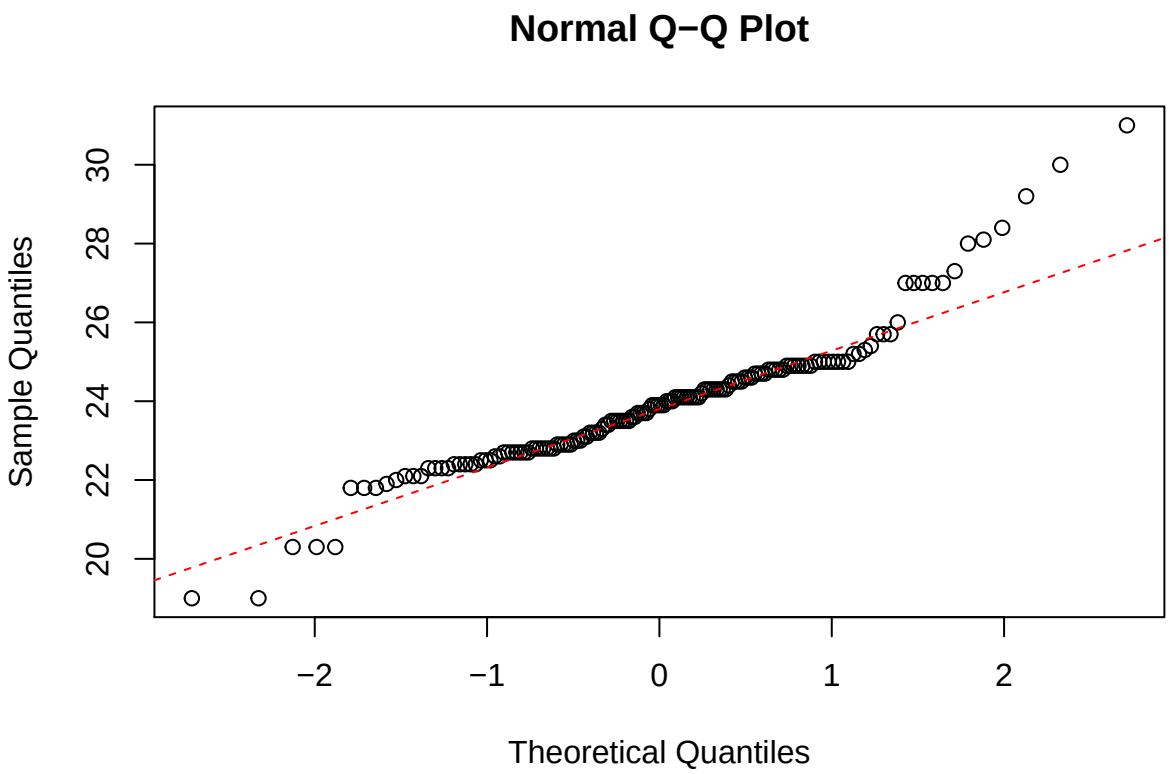
Analysis of the BMI variable

```
# A histogram with the density plot
hist(test$BMI,prob=T)
lines(density(test$BMI),col=2)
```



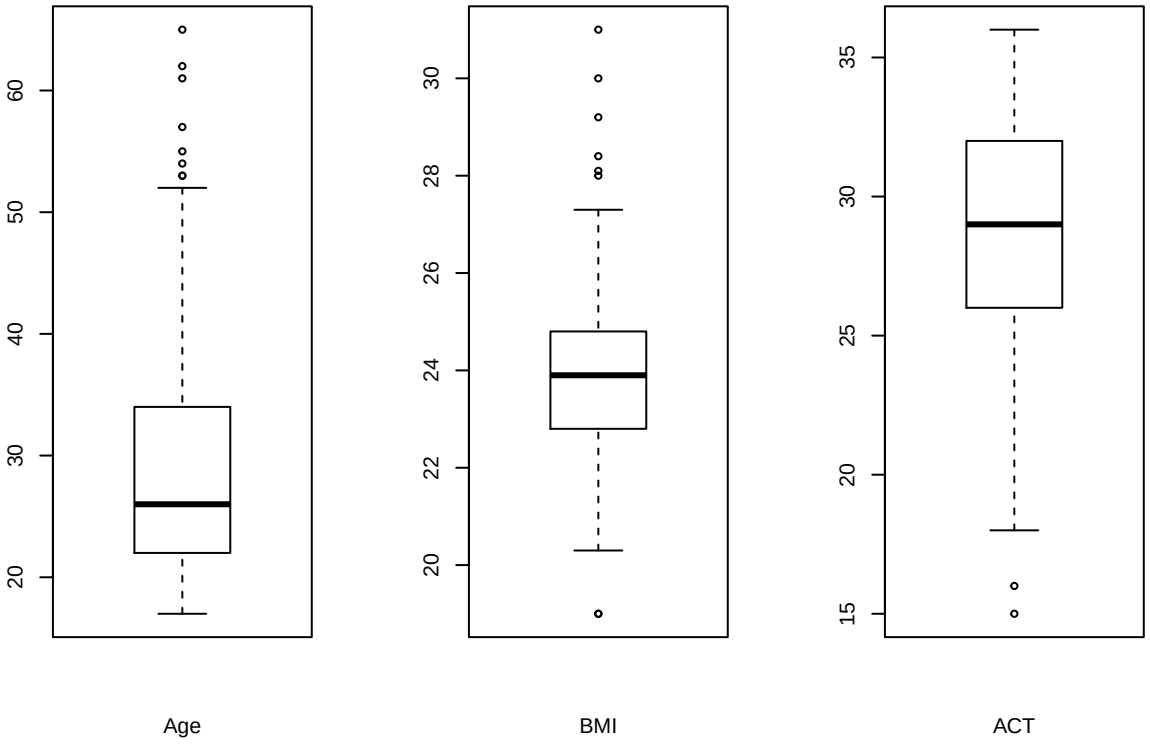
The distribution looks simmetric but there is the presence of outliers (values to low and to high respect to the central cloud)

```
qqnorm(test$BMI)
qqline(test$BMI,col=2,lty=2)
```



The QQplot confirms the presence of anomalous values of BMI.
A unique plot with many boxplots.

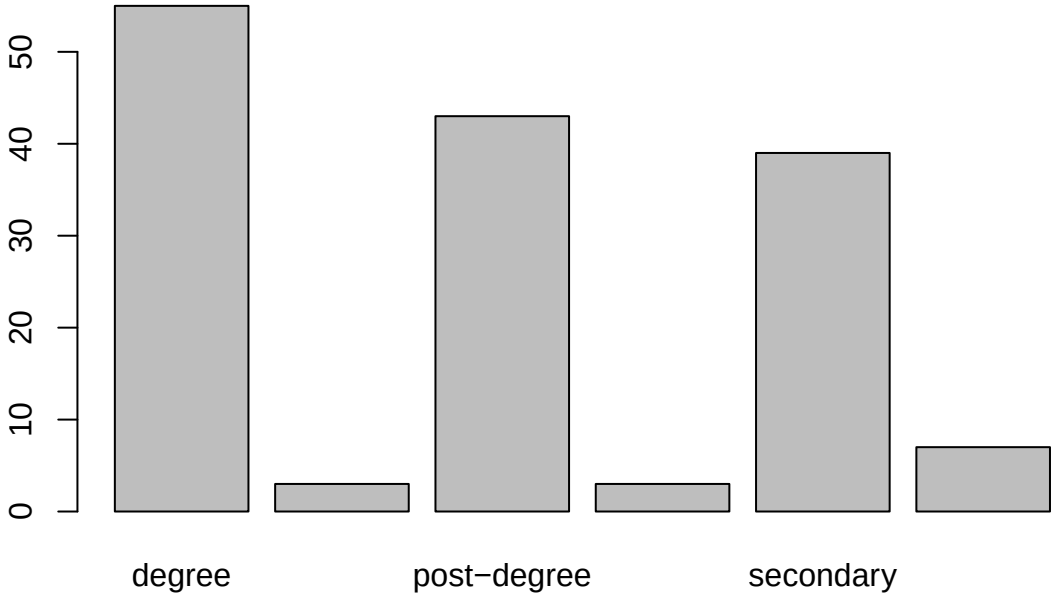
```
par(mfrow=c(1,3)) # 1  row 3 cols
boxplot(test$Age,xlab="Age")
boxplot(test$BMI,xlab="BMI")
boxplot(test$ACT,xlab="ACT")
```



```
par(mfrow=c(1,1))
```

Analysis of the Education variable

```
# A barplot with the density plot
plot(test$Education)
```

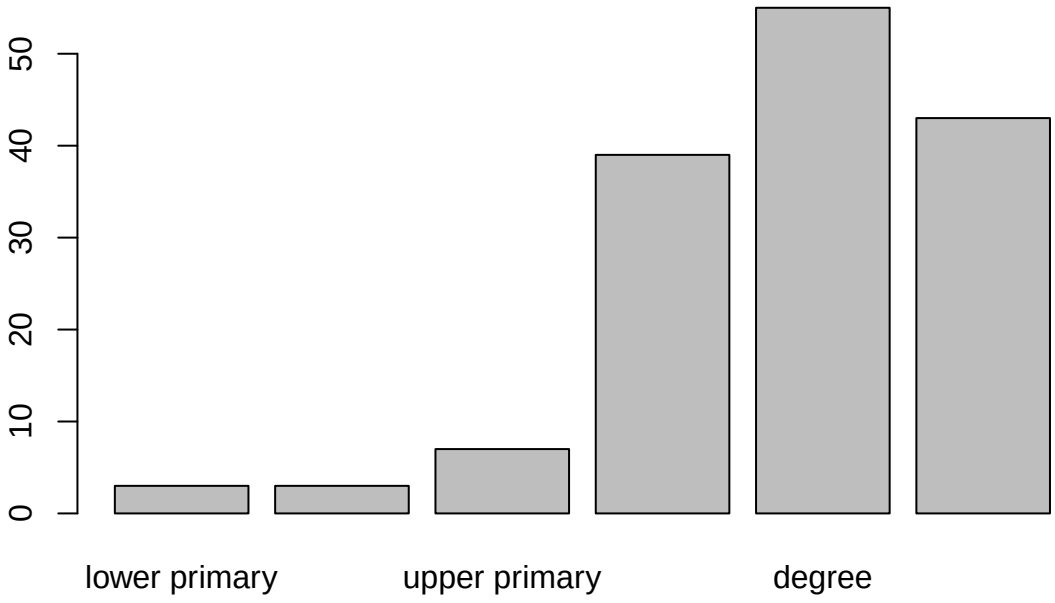


The barplot reports the frequency of each modality of the categorical variable. But in this variable there is an order. So we define the order as follows

```
#here the levels
levels(test$Education)
```

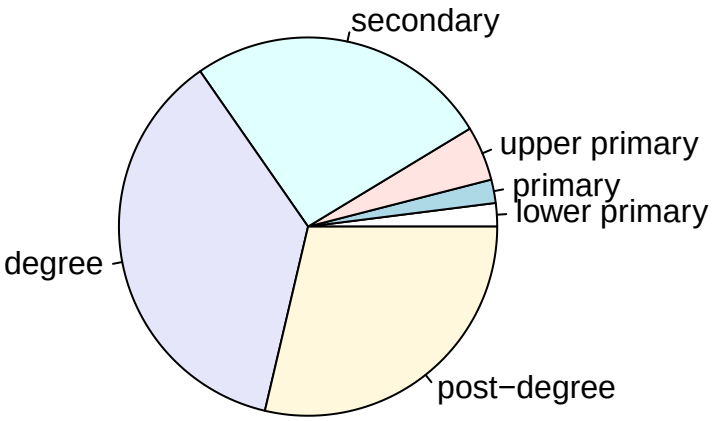
```
## [1] "degree"        "lower primary" "post-degree"   "primary"
## [5] "secondary"     "upper primary"
```

```
test$Education<-factor(test$Education,levels=
c("lower primary","primary","upper primary",
"secondary","degree","post-degree"),ordered =TRUE)
plot(test$Education) # here is ordered
```



Here a pie plot

```
pie(table(test$Education))
```



The function *table()* permits to obtain a frequency table

```
table(test$Education)
```

```
##
## lower primary       primary upper primary     secondary        degree
##             3             3             7            39            55
##   post-degree
##            43
```

```
#or a relative frequency table with the function prop.table()
prop.table(table(test$Education))
```

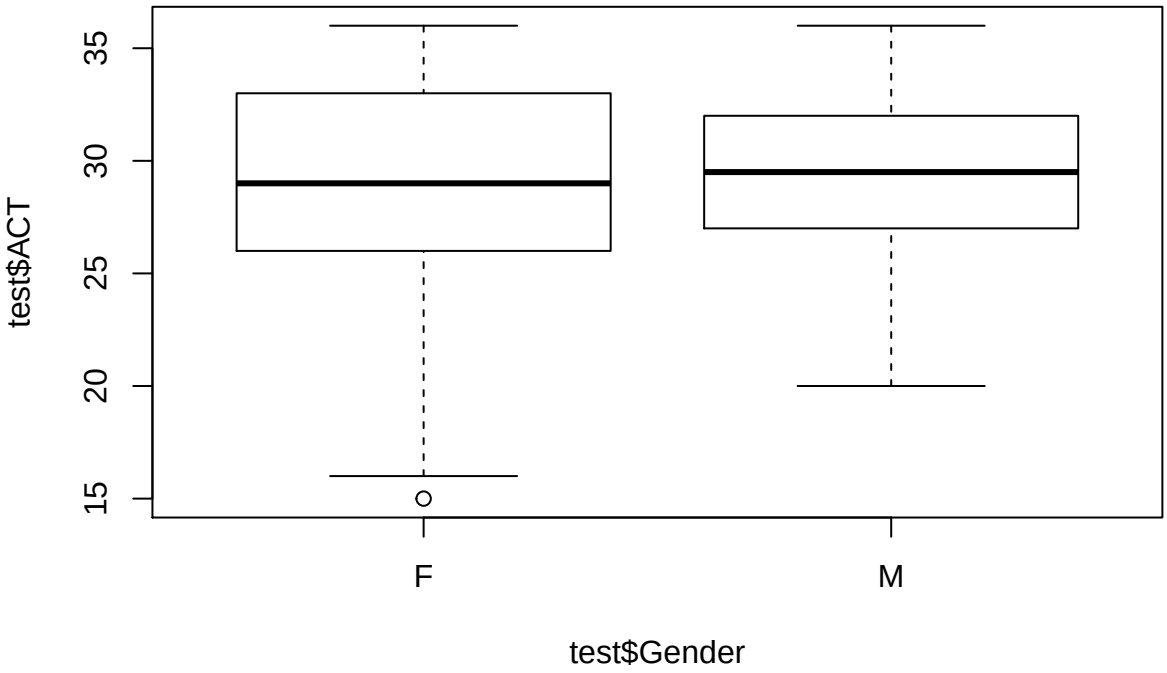
```
##
## lower primary       primary upper primary     secondary        degree
##    0.02000000    0.02000000    0.04666667    0.26000000    0.36666667
##   post-degree
##    0.28666667
```

Bivariate Statistical Analysis with R

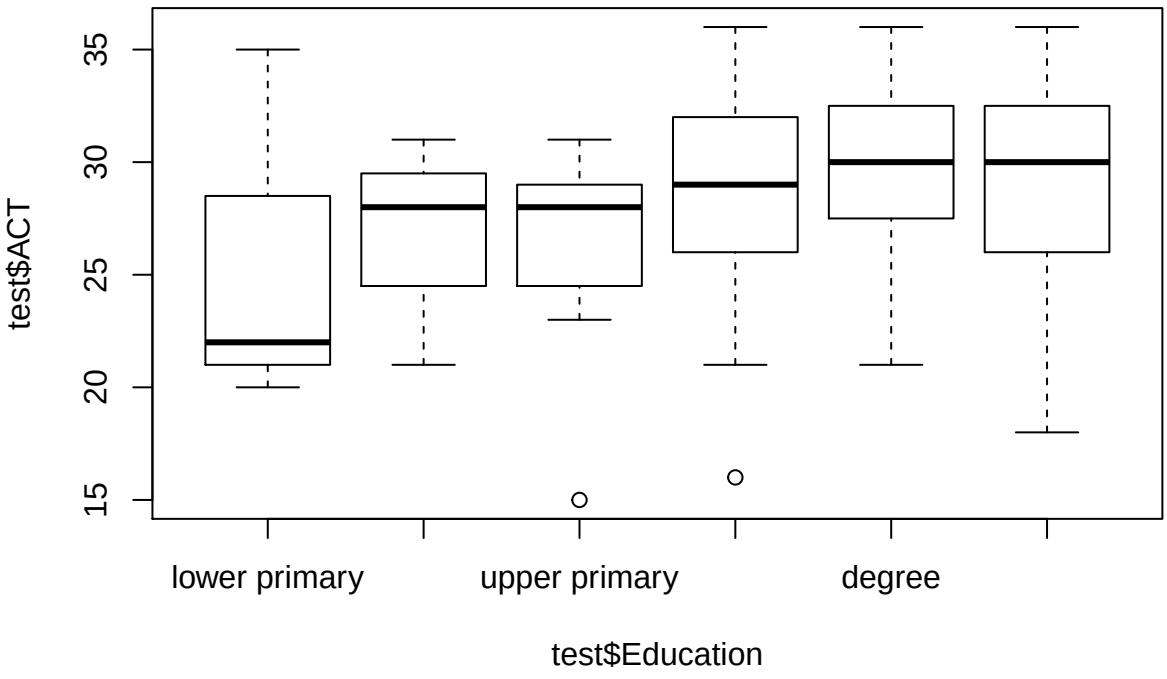
The dataset reported the results of 150 subjects on ACT e SAT tests.
Some variables influences the performances.

Quantitative vs qualitative variables

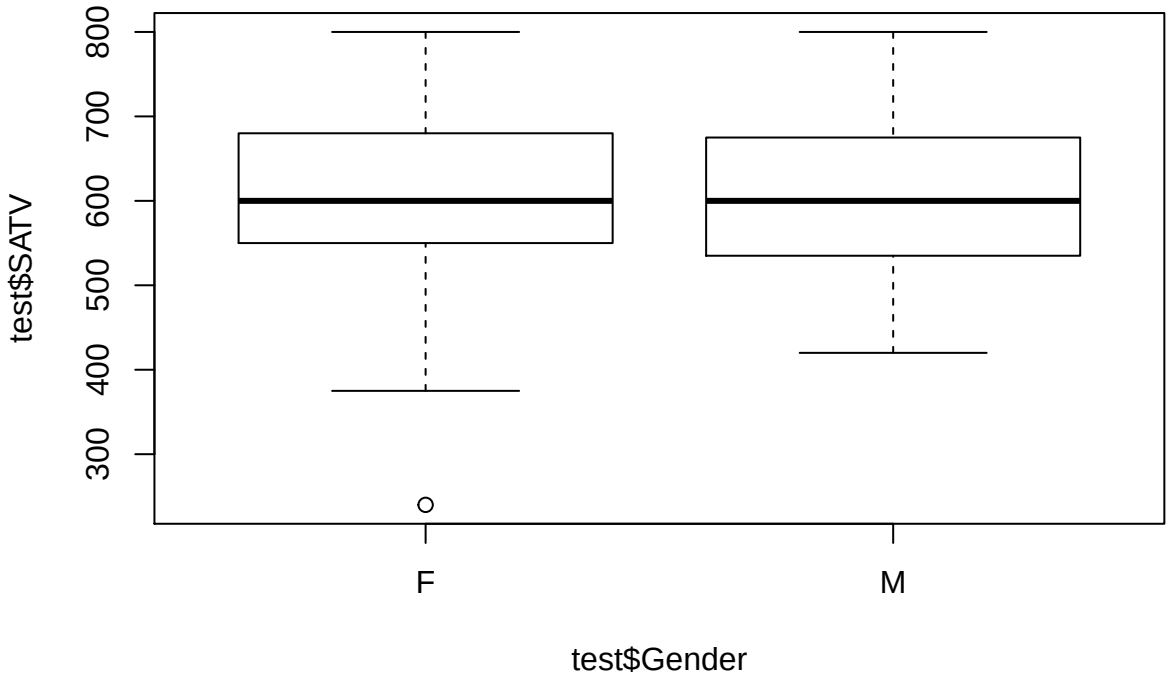
```
#ACT vs Gender and Education
boxplot(test$ACT~test$Gender)
```



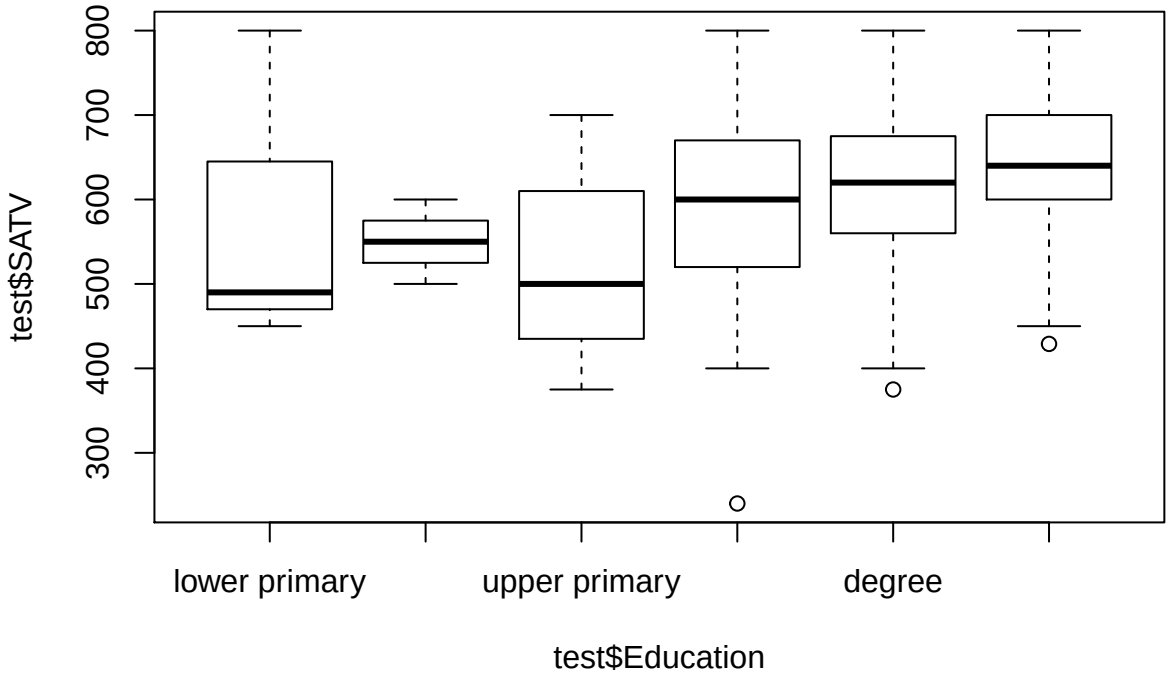

```
boxplot(test$ACT~test$Education)
```



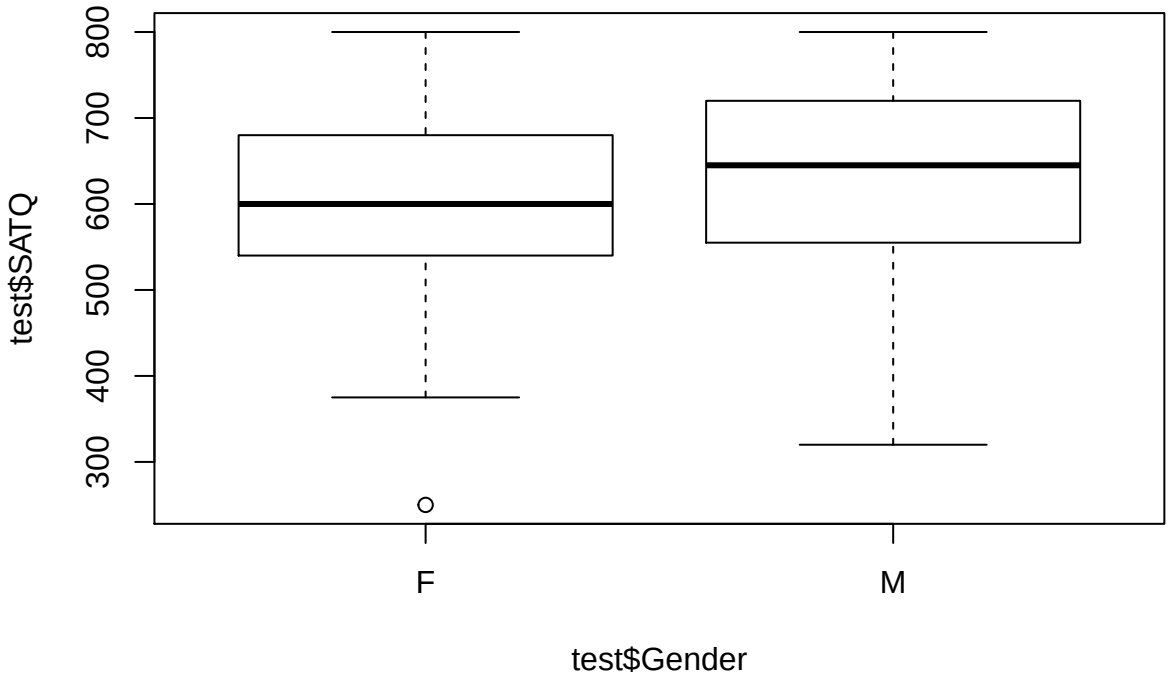
```
#SATV vs Gender and Education
boxplot(test$SATV~test$Gender)
```



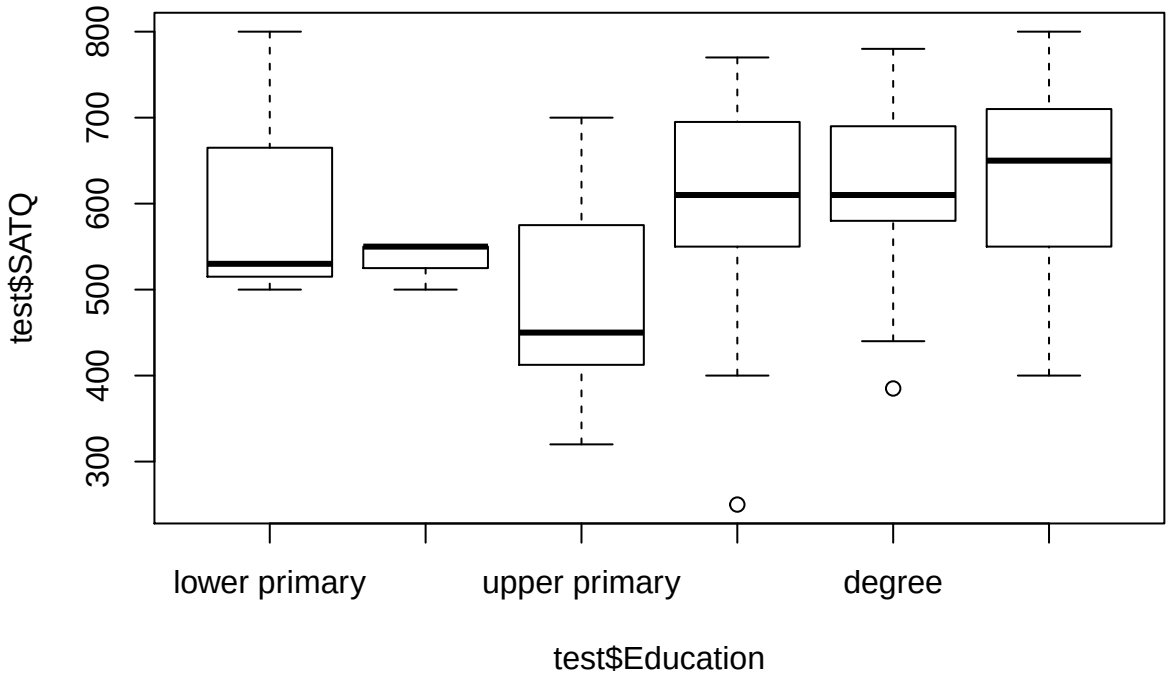

```
boxplot(test$SATV~test$Education)
```




```
#SATQ vs Gender and Education
boxplot(test$SATQ~test$Gender)
```

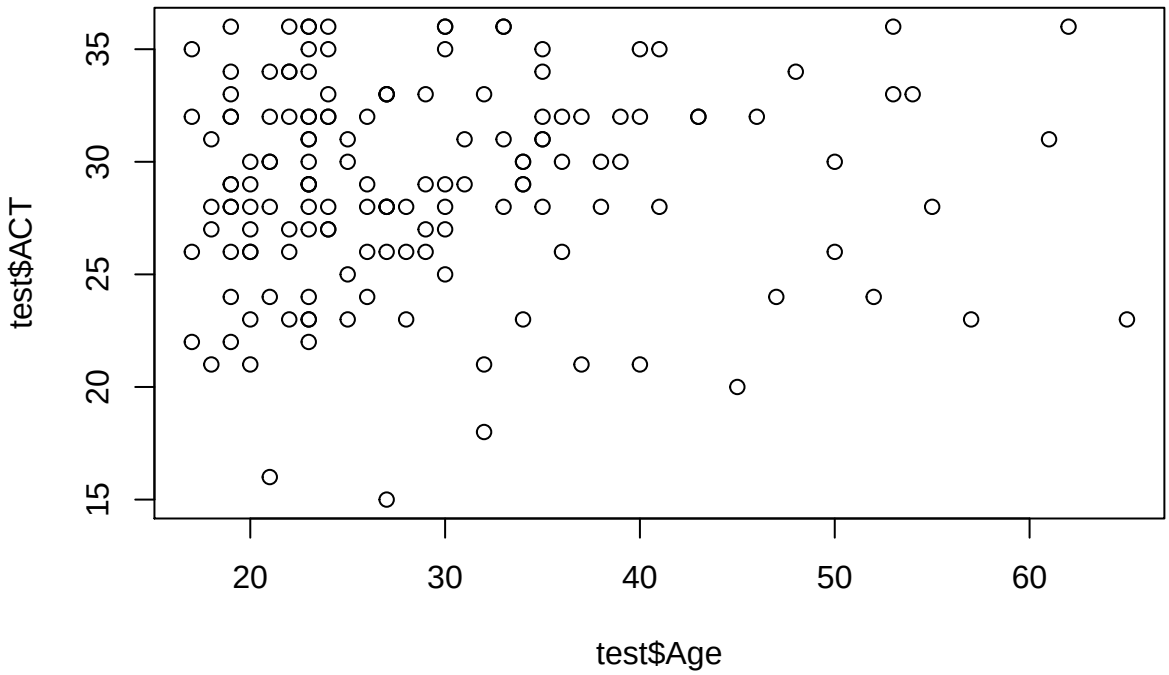



```
boxplot(test$SATQ~test$Education)
```



Quantitative vs Quantitative variables

```
#ACT vs Age
plot(test$Age,test$ACT)
```



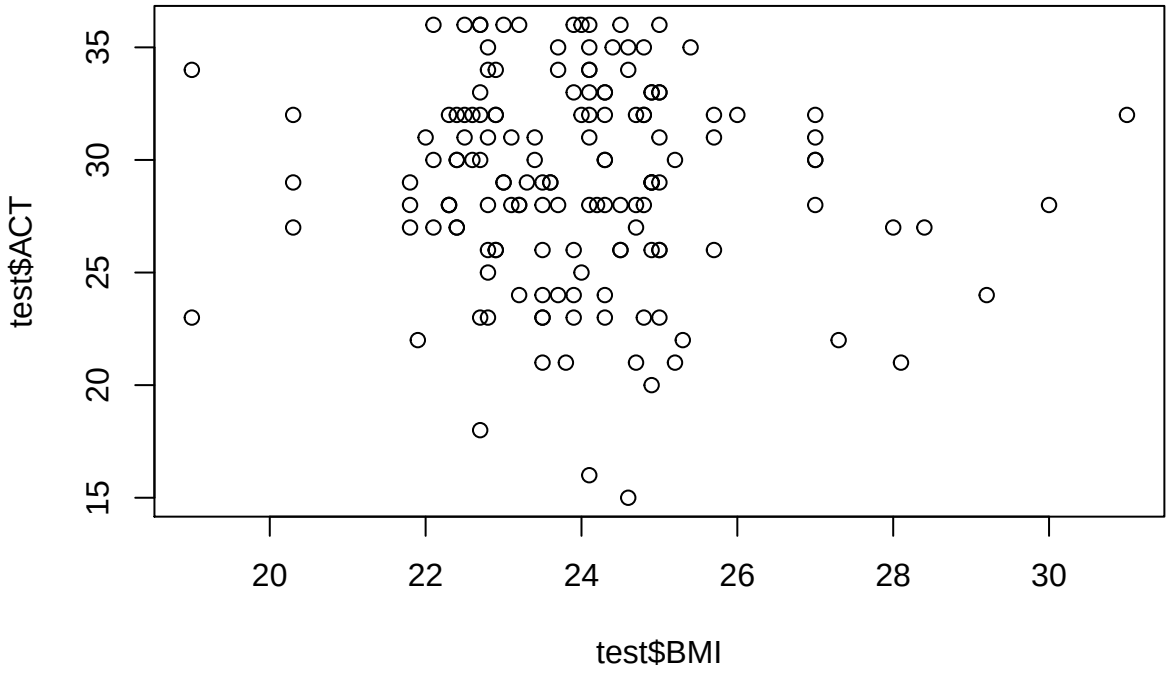

```
cor(test$Age,test$ACT) # pearson
```

```
## [1] 0.06821767
```

```
cor(test$Age,test$ACT,method="spearman") # spearman
```

```
## [1] 0.1033471
```

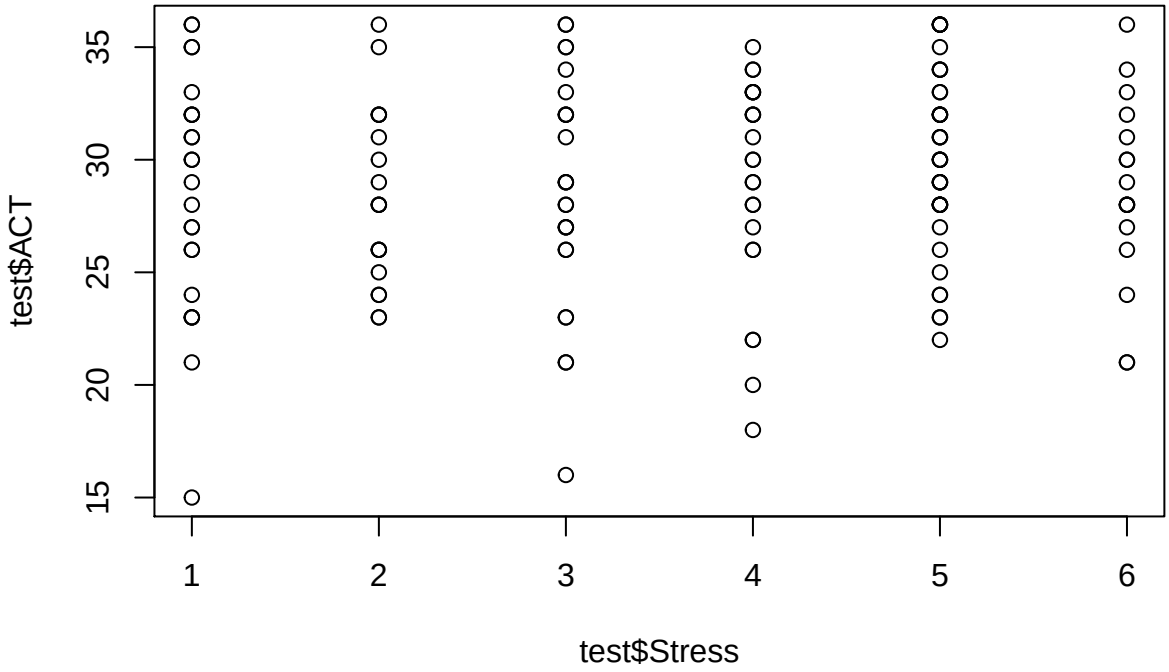
```
#ACT vs BMI
plot(test$BMI,test$ACT)
```




```
cor(test$BMI,test$ACT,method="spearman") # spearman
```

```
## [1] -0.0498391
```

```r
#ACT vs Stress
plot(test$Stress,test$ACT)
```




```r
cor(test$Stress,test$ACT,method="spearman") # spearman
```

```
## [1] 0.08182937
```