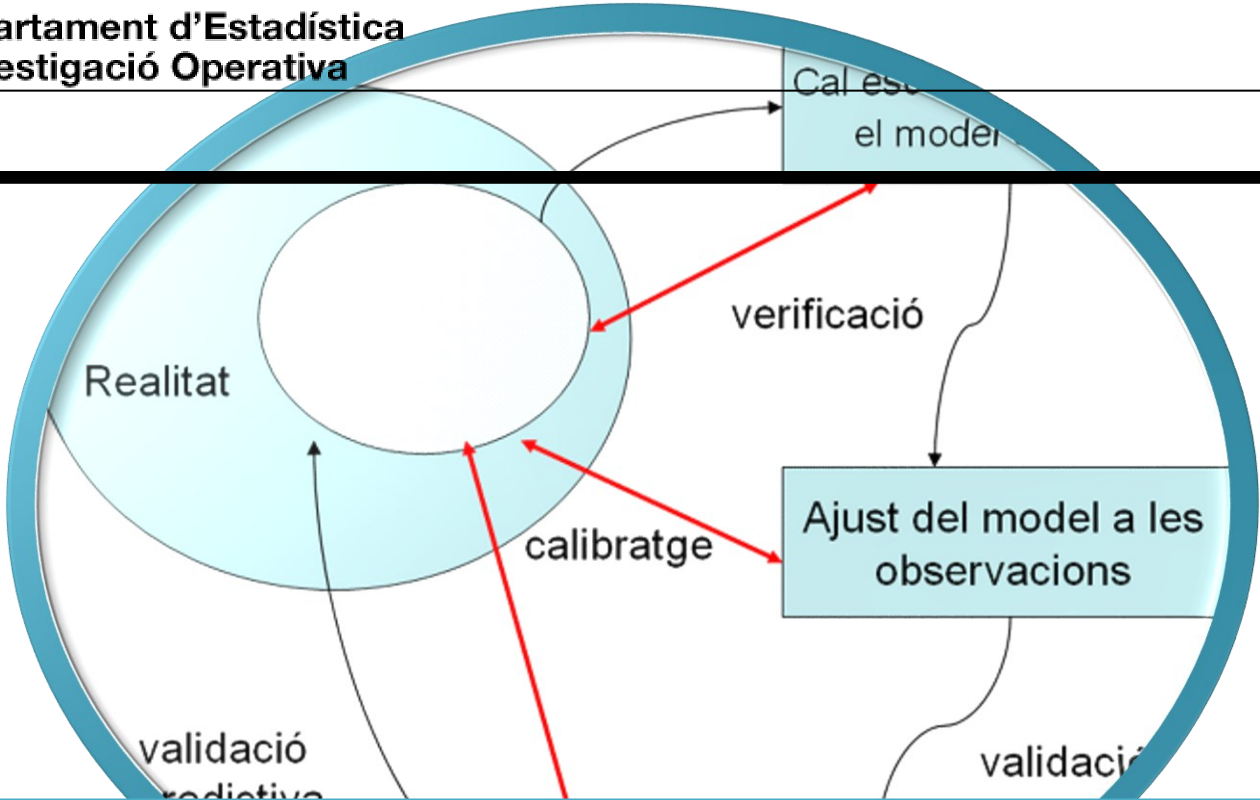
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Estadística
i Investigació Operativa

ACADEMIC YEAR 2025-26



P.01 - Estadística, Optimización y Sistemas – Estadística

Block 1. Exploratory Data Analysis

Lecturer: Lídia Montero

November 2025 – Version 1.1

TABLE OF CONTENTS

2-1. EXPLORATORY DATA ANALYSIS IN R 3

2-1.1 BASIC CONCEPTS OF DESCRIPTIVE DATA ANALYSIS 3
2-1.2 TYPOLOGY OF VARIABLES 4
2-1.3 STATISTICAL PREDICTION MODELS 5
2-1.4 UNIVARIATE DESCRIPTIVE ANALYSIS 6
2-1.4.1 CONTINUOUS UNIVARIATE ANALYSIS DESCRIPTION: NUMERIC INDICATORS 7
2-1.4.2 UNIVARIATE ANALYSIS DESCRIPTION CATEGORICAL 9
2-1.5 BOX-PLOT 14
2-1.6 BIVARIATE DESCRIPTIVE ANALYSIS 15

2-2. EDA IN R – BIVARIATE: 2 NUMERICS Y VS X 18

2-2.1 NUMERIC STATISTICS TO ASSESS LINEAR RELATIONSHIP BETWEEN Y AND X 18

2-3. EDA IN R – BIVARIATE: 2 NUMERICS Y VS X 19

2-4. EDA IN R – BIVARIATE: 2 FACTORS, A AND B 20

2-4.1 NUMERIC STATISTICS TO ASSESS LINEAR RELATIONSHIP A AND B 20

2-1. EXPLORATORY DATA ANALYSIS IN R

2-1.1 Basic concepts of descriptive data analysis

Data matrix structure (data.frame in R)

POPULATION

	Carac1	Carac2	…
Individual 1	value	value’	…
Individual 2	value’’	value’’’	…
.			
.			

```
   edad        residencia llista.ED llista.ED_1
1    19           BCN-AMB        22          22
2    20           BCN-AMB        25          25
3    19           BCN-AMB        34          34
4    20           BCN-AMB        35          35
5    19           BCN-AMB        41          41
6    19           BCN-AMB        41          41
7     9           BCN-AMB        46          46
8    20           BCN-AMB        46          46
9    19           BCN-AMB        46          46
10   19           BCN-AMB        47          47
11   19 Resta Catalunya        49          49
12   19 Resta Catalunya        54          54
13   23 Resta Catalunya        54          54
14   19   Estat Espanyol        59          59
15   19   Estat Espanyol        60          60
16   NA              <NA>        NA         100
```

Sample: Subset of a population
Features ≡Variables
Values : numeric or alphanumeric

Example 2.1: Age data and residence’s place of students of a UPC’s class

2-1 EXPLORATORY DATA ANALYSIS IN R

2-1.2 Typology of variables

Numerical (continuous)

COVARIATES/ COVARIANTS

Continuous (reals values or simply many different values)

Ex: Incomes, weight, lung capacity, etc.

Discretes (equivalent to whole numbers or natural ... if there are many val e continuous)

Ex: Children's number, age, etc.

Categorical (cualitatives)

FACTORS

(values : modalities or categories)

<u>With order (ordinal)</u>

Ex: Level of education, Labor category, etc.

<u>Unordered (nominal)</u>

Ex: Gender, Race, Marital status ...

Categorical variables come pruned expressed by a numerical value (Ex. Gender: Man = 0, Woman = 1). (Not to be confused with quantitative variables)

2-1 EXPLORATORY DATA ANALYSIS IN R

2-1.3 Statistical Prediction Models

- Interest: explain one (or more) response variable or dependent.
- From explanatory variables or predictors.

Classification of variables:

- Pure nominal or categorical variables: binary (dichotomous) if they have 2 categories and polytomous if they have more than 2 categories. The categories do not have any semantics associated order. They are qualitative variables.
- Ordinal Variables. They are categorical variables with notion of order among the categories, usually more than 2. ***They often come from the discretization of continuous variables*** or are discrete a.v.. They are qualitative variables.
- Continuous or quantitative variables. Theoretically associated with continuous measures.
- Factor: qualitative variable explanatory. The different categories are called levels.
- Covariant: continuous explanatory variable.

2-1 EXPLORATORY DATA ANALYSIS IN R

2-1.4 Univariate descriptive analysis

Continuous variable description: *Missing* and *Outliers*

- **Numerical values**
  - Measures of Central Tendency: *Mean, Median, Mode*
  - Measures of Dispersion: *Variance, Standard Deviation, Quartiles, IQR, Maximum, Minimum.*
- **Graph Representations**
  - Histogram, Cumulative Histogram. Absolute or relative.
  - *BoxPlot*.

Description of a categorical variable: Graph Representations

- Bar chart: absolute or relative.
- *Pie Chart*.

2-1 EXPLORATORY DATA ANALYSIS IN R

2-1.4.1 Continuous Univariate Analysis Description: Numeric Indicators

```
> summary(dataframe)
```

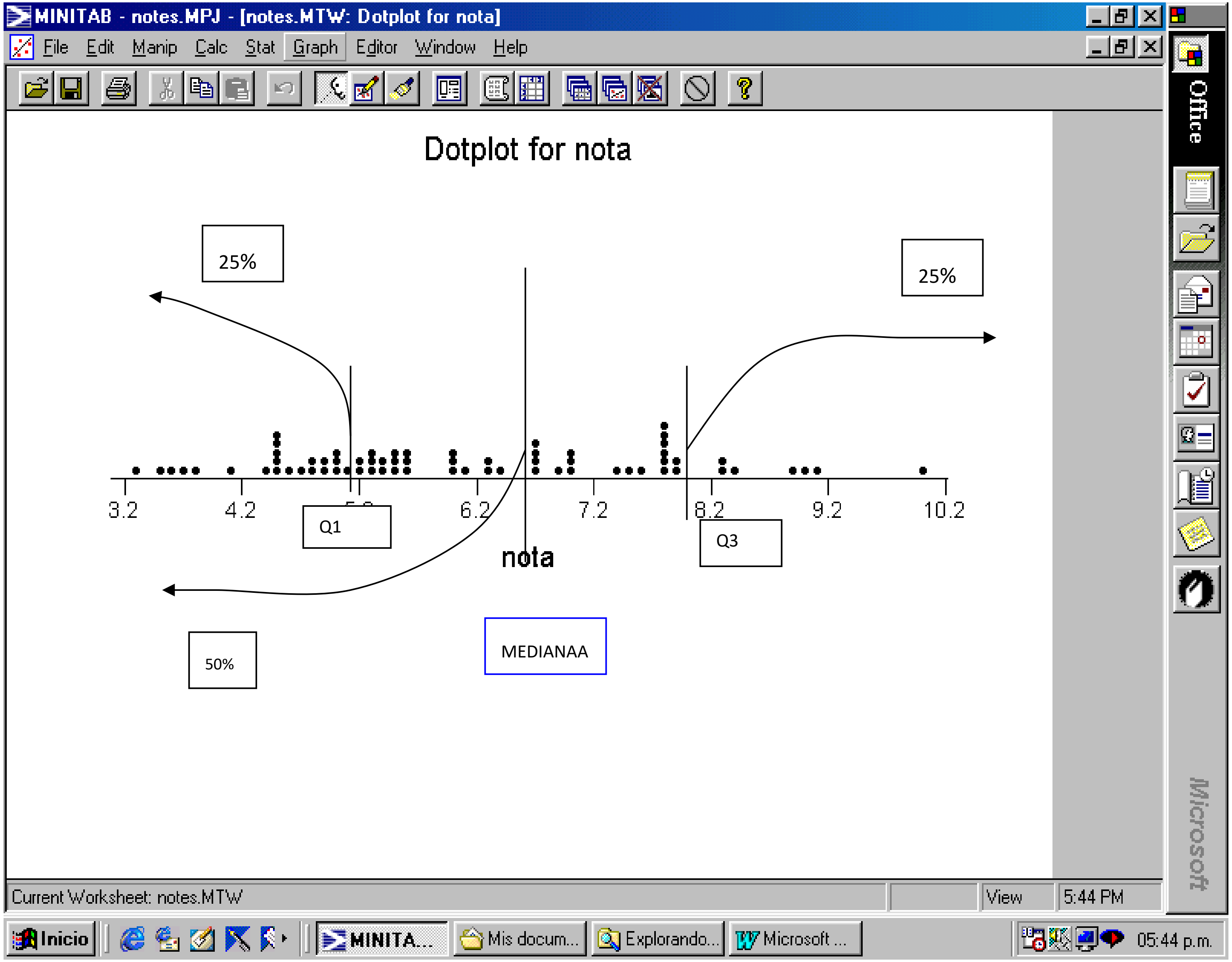
- Mean $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Median: Value of the *variable* such that

  50% Observations are < Median (Q2) & 50% Observations are > Median (Q2)
- Quartile Q1 of the 25% and quartile Q3 of the 75%: Values of the variable that

  25% Observations are < Q1 & 75% Observations are > Q1

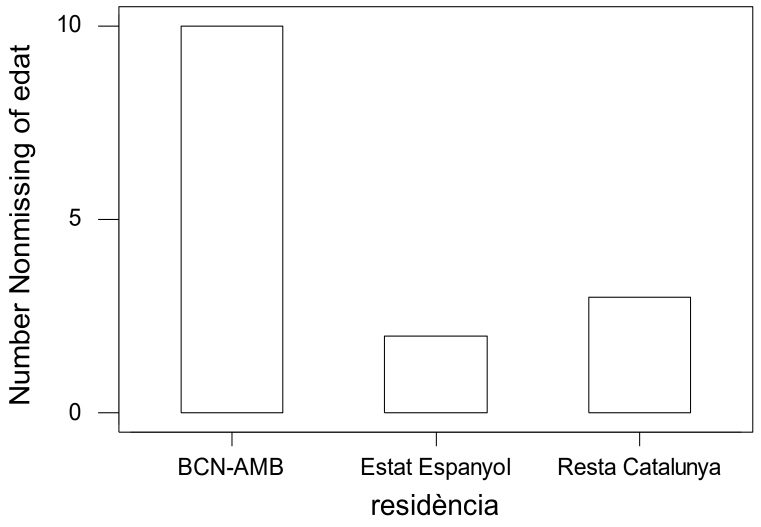
  75% Observations are < Q3 & 25% Observations are > Q3
- Variance $\boldsymbol{s_x}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$
- Standard Deviation $\boldsymbol{s_x}$

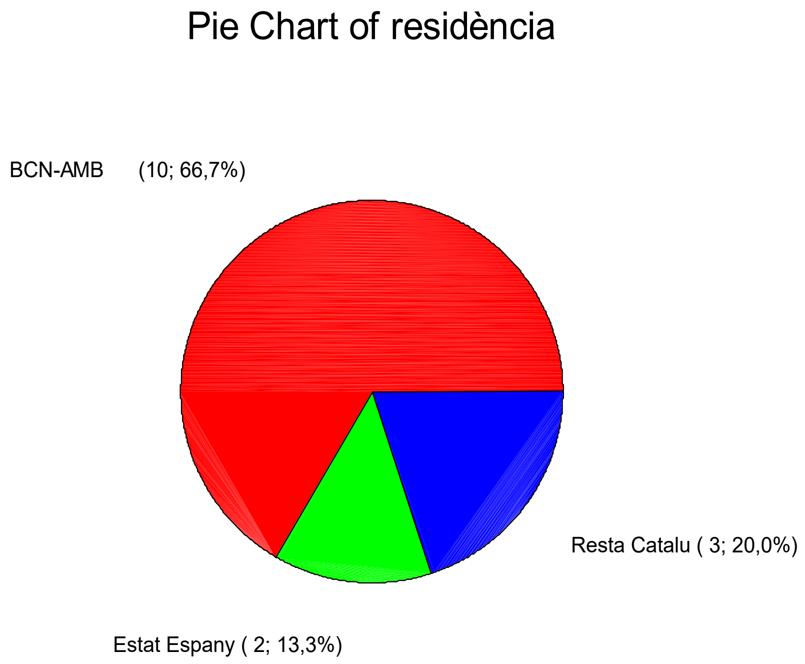
2-1 EXPLORATORY DATA ANALYSIS IN R



2-1 EXPLORATORY DATA ANALYSIS IN R

2-1.4.2 Univariate Analysis Description categorical





Bar chart (absolute or relative)

$\boxed{\textbf{barplot(table()) in R}}$

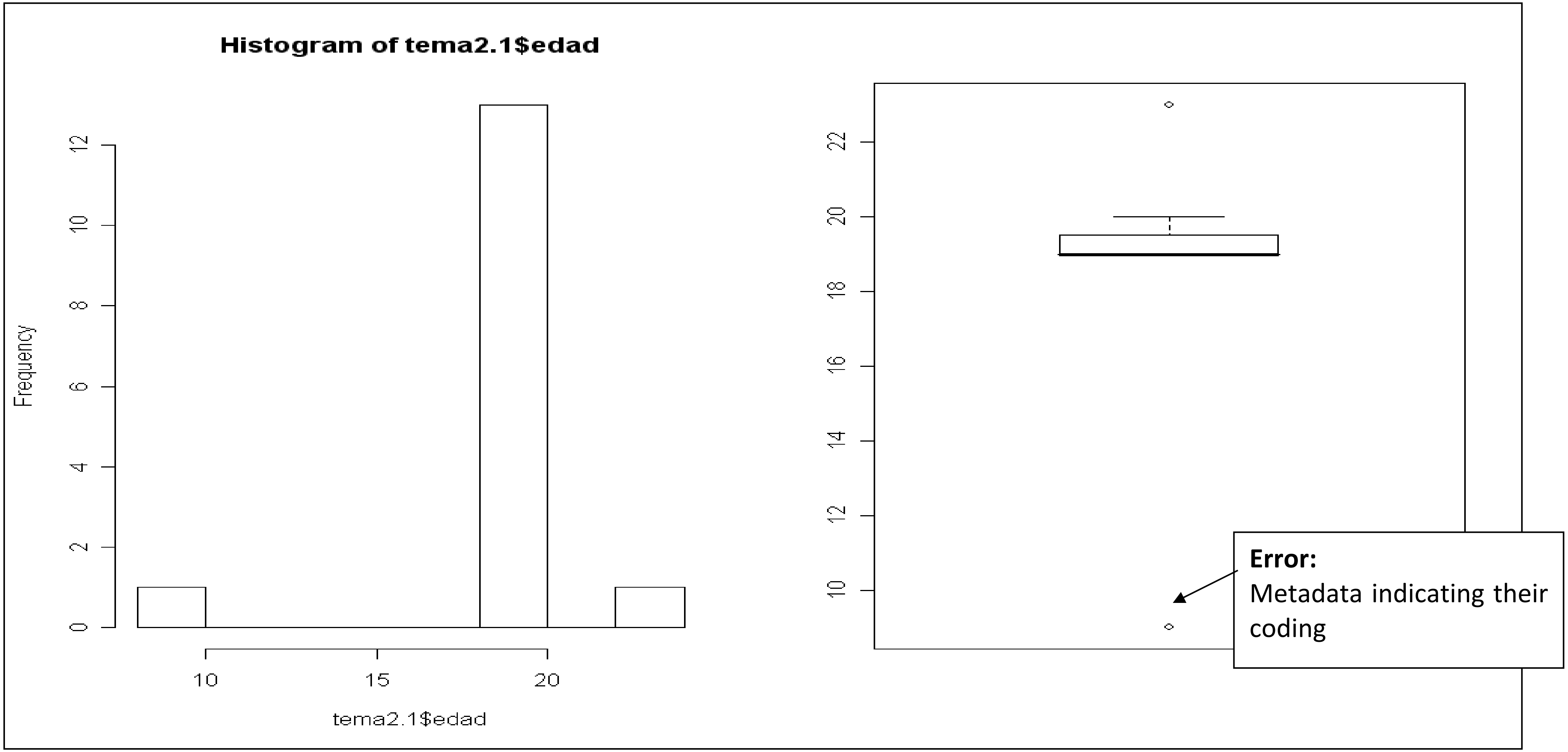
Pie Chart

2-1 EXPLORATORY DATA ANALYSIS IN R

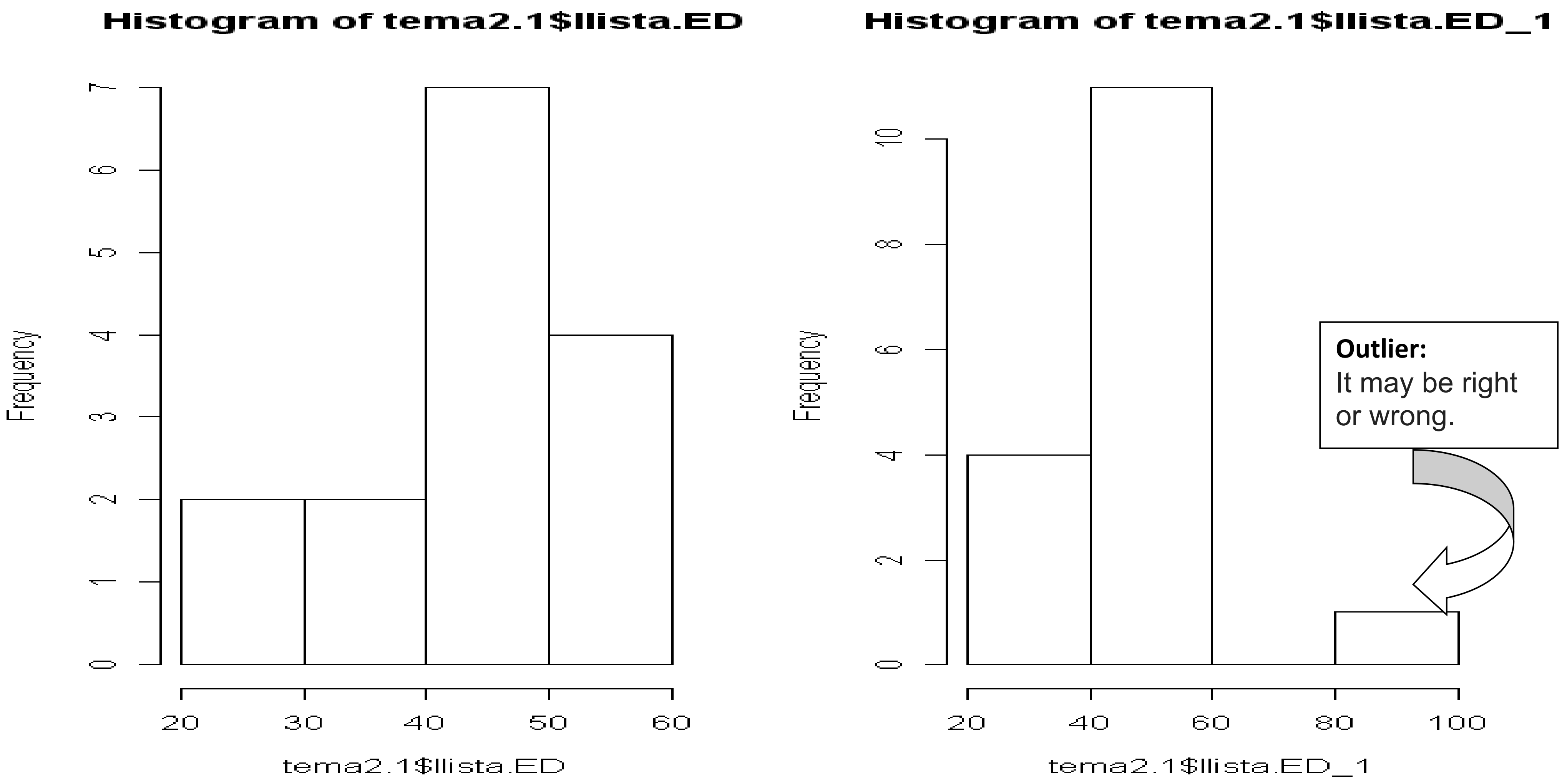
```
> tema2.1 <- read.table("tema2.1.txt",header=T,sep='\t',na.string='')
> tema2.1
   edad      residencia llista.ED llista.ED_1
1    19         BCN-AMB        22          22
2    20         BCN-AMB        25          25
3    19         BCN-AMB        34          34
4    20         BCN-AMB        35          35
5    19         BCN-AMB        41          41
6    19         BCN-AMB        41          41
7     9         BCN-AMB        46          46
8    20         BCN-AMB        46          46
9    19         BCN-AMB        46          46
10   19         BCN-AMB        47          47
11   19 Resta Catalunya        49          49
12   19 Resta Catalunya        54          54
13   23 Resta Catalunya        54          54
14   19  Estat Espanyol        59          59
15   19  Estat Espanyol        60          60
16   NA            <NA>        NA         100
> summary(tema2.1)
      edad                 residencia   llista.ED      llista.ED_1
 Min.   : 9.0   BCN-AMB        :10   Min.   :22.00   Min.   : 22.00
 1st Qu.:19.0   Estat Espanyol : 2   1st Qu.:38.00   1st Qu.: 39.50
 Median :19.0   Resta Catalunya: 3   Median :46.00   Median : 46.00
 Mean   :18.8   NA's           : 1   Mean   :43.93   Mean   : 47.44
 3rd Qu.:19.5                        3rd Qu.:51.50   3rd Qu.: 54.00
 Max.   :23.0                        Max.   :60.00   Max.   :100.00
 NA's   : 1.0                        NA's   : 1.00
```

2-1 EXPLORATORY DATA ANALYSIS IN R

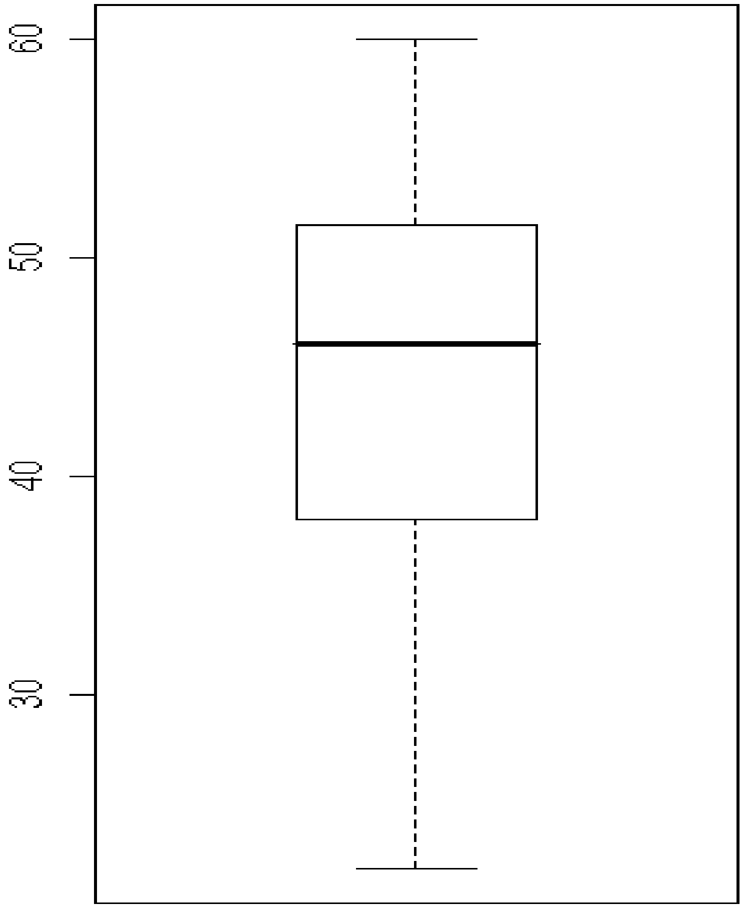
```
par(mfrow=c(1,2))
hist(tema2.1$edad)
boxplot(tema2.1$edad)
```

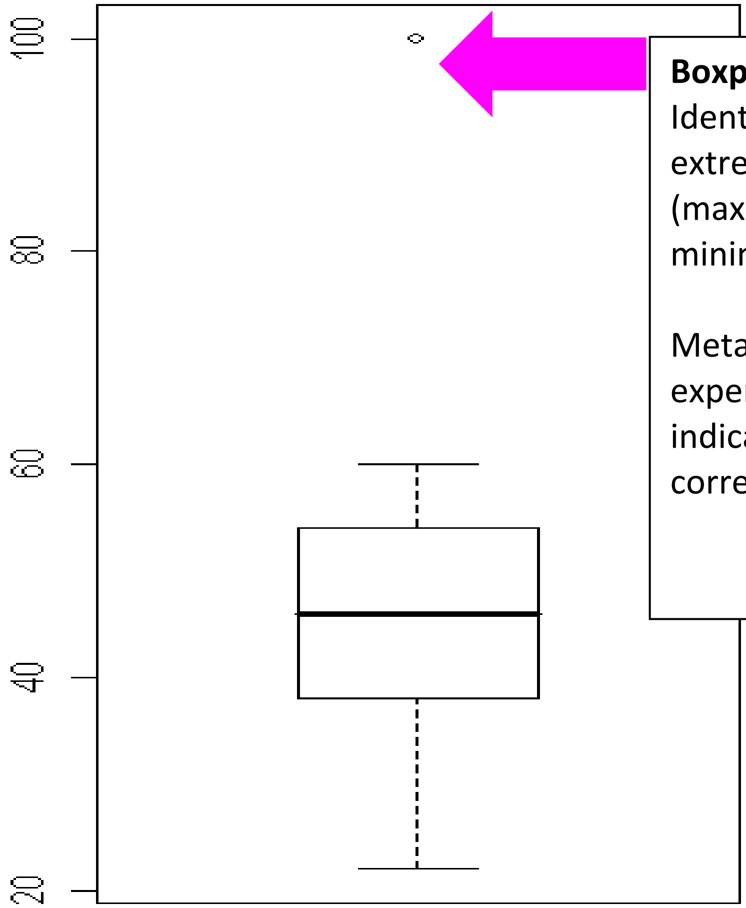


2-1 EXPLORATORY DATA ANALYSIS IN R



2-1 EXPLORATORY DATA ANALYSIS IN R





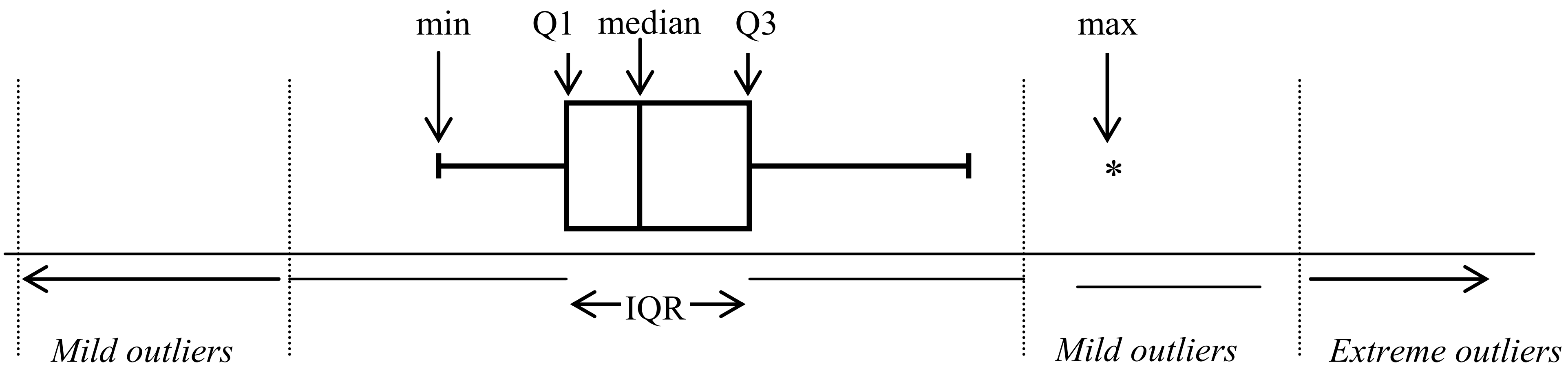
Boxplot:
Identification of extreme values (maximum and minimum).

Metadata or experience indicates if they are correct or not.

2-1 EXPLORATORY DATA ANALYSIS IN R

2-1.5 Box-plot

"Five issues Summary" (Min, Q1, Me, Q3, Max) for Univariate DE to detect the existence of outliers.



The area between Q3 and Q3+1.5 IQR and Q3+3IQR is called mild outliers upper zone. Similarly with the lower tail: between Q1-1,5IQR and Q1-3IQR. The area above the point Q3+3IQR area called extreme outliers. As a general rule, it isn't worrying to see up to 1% of extreme outliers and up to 5% of mild outliers in any distribution.

2-1 EXPLORATORY DATA ANALYSIS IN R

2-1.6 Bivariate descriptive analysis

Study of the relationship between variables in pairs. Naturally, is the simplest case of multivariate descriptive analysis, that globally study the relationships among a set of variables that can be very large (more complex techniques that connect directly with Data Mining).

The most common techniques of bivariate descriptive analysis, as happened in the univariate case, are of two types:

- Graph: Allow display as the relationship between two variables.
- Numeric: Quantify what you see on the graph with a appropriate statistic.

The nature of the variables to study plays a key role in determining the tools to use in each case. Three cases are distinguished primarily:

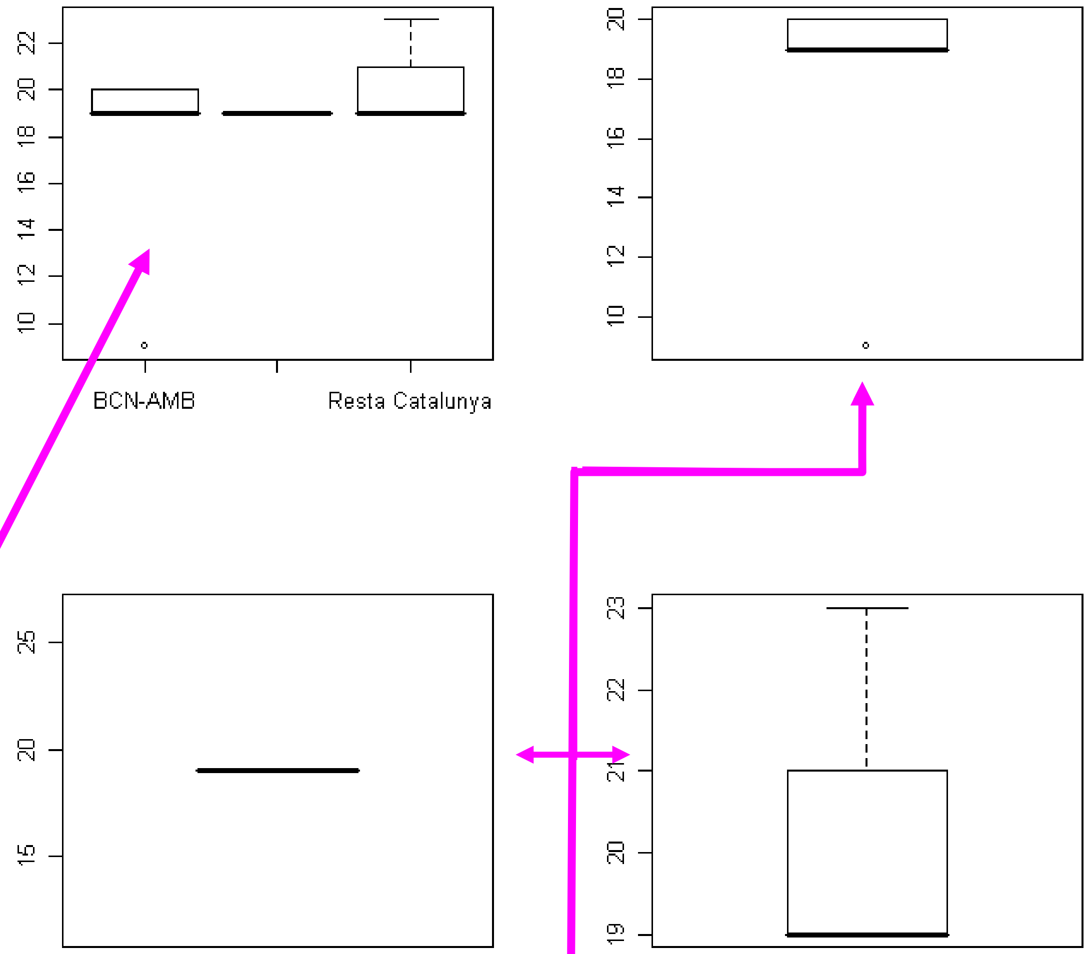
- Relationships between a numeric variable and a categorical. For example, descriptive groups.
- Relationships between two categorical variables. For example, contingency tables.
- Relationships between two quantitative variables. For example, simple linear regression.

2-1 EXPLORATORY DATA ANALYSIS IN R

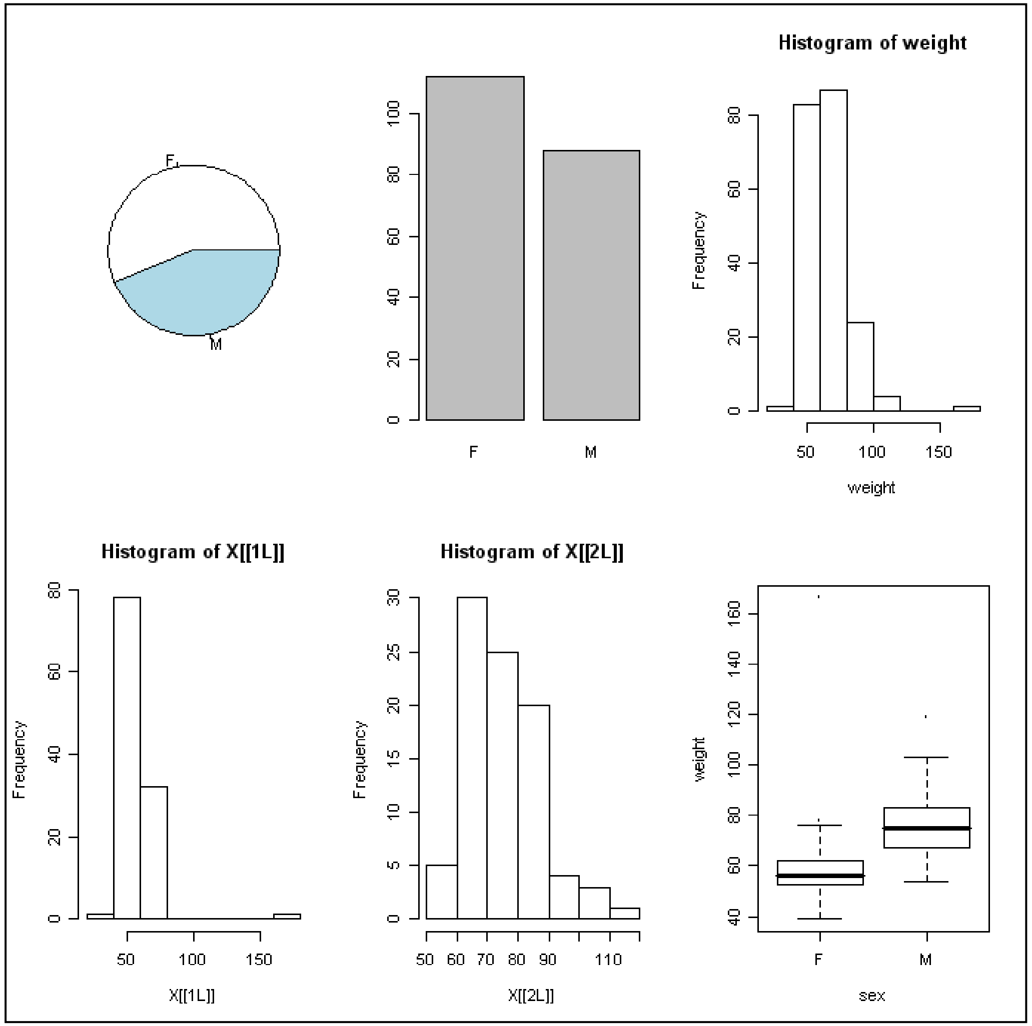
In example you can try a descriptive groups, consider age as a response variable and place of residence as the explanatory variable.

```
# AD Bivariant per grups
>tapply(tema2.1$edad,tema2.1$residencia,mean)
BCN-AMB  Estat Espanyol Resta Catalunya
18.30000        19.00000        20.33333

tapply(tema2.1$edad,tema2.1$residencia,summary)
"BCN-AMB"
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  9.00   19.00   19.00   18.30   19.75   20.00
"Estat Espanyol"
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    19      19      19      19      19      19
"Resta Catalunya"
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 19.00   19.00   19.00   20.33   21.00   23.00
> attach(tema2.1)
> par(mfrow=c(2,2))
> plot(edad~residencia,data=tema2.1)
>
apply(tema2.1$edad,tema2.1$residencia,boxplot)
```



2-1 EXPLORATORY DATA ANALYSIS IN R – BIVARIATE: NUMERIC VS FACTOR




```
par(mfrow=c(2,3))
attach(Davis)
pie( table( sex ))
barplot( table(sex) )
hist( weight )
```

```
tapply( weight, sex, hist )# Not nice
plot( weight ~ sex ) # Boxplot is
default plot
```

2-2. EDA IN R – BIVARIATE: 2 NUMERICS Y VS X

2-2.1 Numeric statistics to assess linear relationship between Y and X

Covariance, COV(y,x)=COV(x,y), defined as E(YX) – E(X)E(Y)

- Disadvantage: Depends on units, so not direct interpretation

Pearson's coefficient of correlation, suitable for assessment in normal data

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \quad and \quad \sigma_X = \sqrt{Var(X)} \; \sigma_Y = \sqrt{Var(Y)}$$

- Advantage: Adimensional, no affected by units
  - $\rho(X, Y)$ **range is** $[-1,1]$.
    - $\rho(X, Y)$ > 0 means positive relationship X and Y.
    - $\rho(X, Y)$ < 0 means negative relationship X and Y,.
    - $\rho(X, Y)$ = 0 indicates uncorrelated variables, not equivalent to independence.
  - If Y = aX + b then $|\rho(X, Y)| = 1$.
- **Spearman's coefficient of correlation,** is a nonparametric measure of statistical dependence.

2-3. EDA IN R – BIVARIATE: 2 NUMERICS Y VS X

In R, use var(Davis[,2:3]) or try with Census Data data("CPS1985") in library AER.

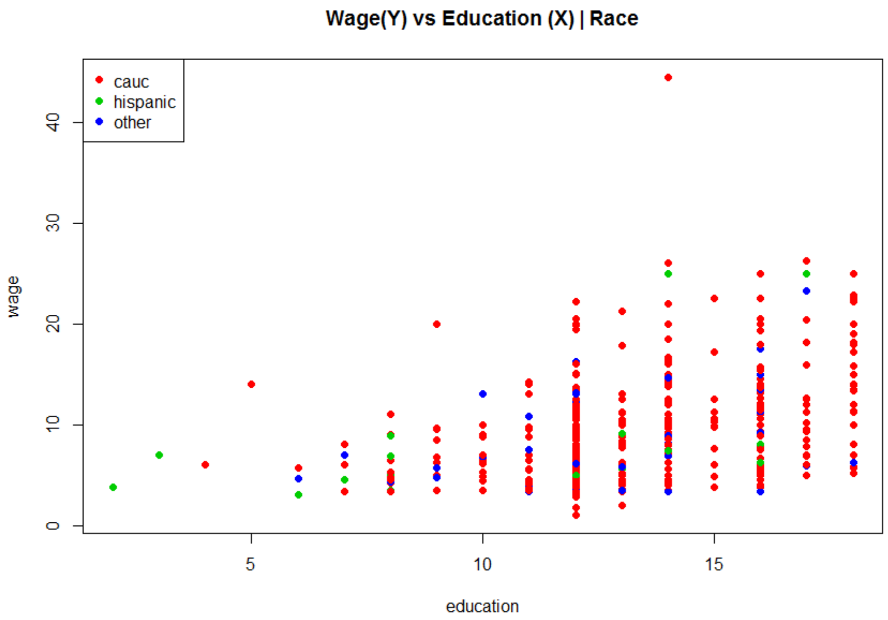
```
> library(AER)
> data("CPS1985")
> df<-CPS1985
> ls()
[1] "CPS1985" "df"
> dim( df )  # dimensions: rows and columns
[1] 534  11
> summary( df )
      wage          education       experience         age          ethnicity      region       gender          occupation
 Min.   : 1.000   Min.   : 2.00   Min.   : 0.00   Min.   :18.00   cauc    :440   south:156   male  :289   worker    :156
 1st Qu.: 5.250   1st Qu.:12.00   1st Qu.: 8.00   1st Qu.:28.00   hispanic: 27   other:378   female:245   technical :105
 Median : 7.780   Median :12.00   Median :15.00   Median :35.00   other   : 67                            services  : 83
 Mean   : 9.024   Mean   :13.02   Mean   :17.82   Mean   :36.83                                           office    : 97
 3rd Qu.:11.250   3rd Qu.:15.00   3rd Qu.:26.00   3rd Qu.:44.00                                           sales     : 38
Max.   :44.500   Max.   :18.00   Max.   :55.00   Max.   :64.00
          sector     union      married
 manufacturing: 99   no :438   no :184
 construction : 24   yes: 96   yes:350
 other        :411


> attach( df )
> # Bivariate analysis: 2 numeric variables
> plot(education,wage,col=as.numeric(ethnicity)+1,
      main="Wage(Y) vs Education (X) | Race",pch=19)

> legend("topleft",legend=levels(ethnicity),col=2:4,
      pch=19)
> cor(wage,education,method="spearman")
[1] 0.3813425
> cor(wage,education,method="pearson")  # The one defined p
[1] 0.3819221


Nicer option: scatterplot, try in lab session

> library(car)
> scatterplot(wage~education|ethnicity,main="Wage(Y) vs Education (X) | Race",smooth=FALSE)
```



2-4. EDA IN R – BIVARIATE: 2 FACTORS, A AND B

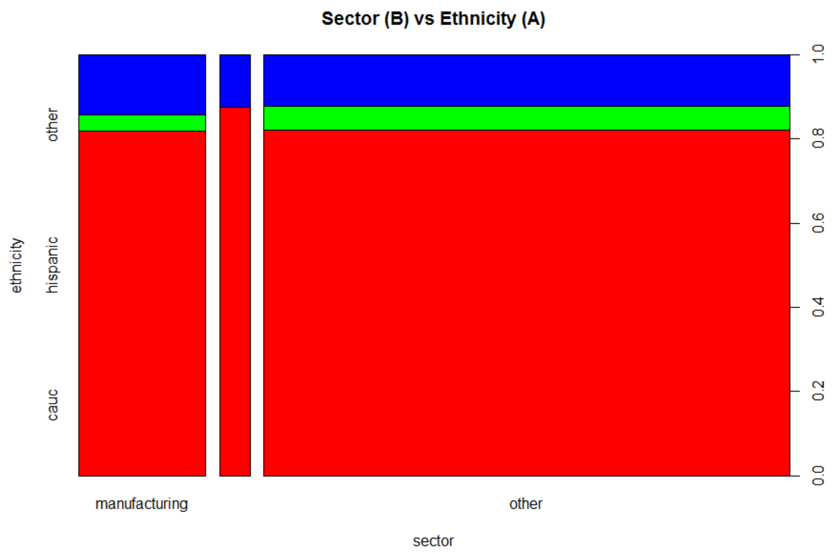
2-4.1 Numeric statistics to assess linear relationship A and B

Non-existent. Analysis of Contingency Tables and classical inference test to assess Independence of both factors using Chi-Squared Test: chisq.test() in R, arguments a contingency table.

```
> ta<-table(ethnicity,sector)
> ta
          sector
ethnicity  manufacturing construction other
  cauc                81           21   338
  hispanic             4            0    23
  other               14            3    50
> round(prop.table(ta,2),2)
          sector
ethnicity  manufacturing construction other
  cauc              0.82         0.88  0.82
  hispanic          0.04         0.00  0.06
  other             0.14         0.12  0.12

> plot(ethnicity~sector,main="Sector (B) vs Ethn
icity (A)",col=rainbow(3))
> chisq.test(ta)

      Pearson's Chi-squared test data:  ta
X-squared = 1.9819, df = 4, p-value = 0.7391
Warning message:In chisq.test(ta) : Chi-squared approximation may be incorrect
```



EDA IN R – BIVARIATE: 2 FACTORS, A AND B

Graphic display (default in R): mosaic plot

More than 2 dimensions: use xtabs() command in R

```
> xtabs(~gender+ethnicity+sector)
, , sector = manufacturing

        ethnicity
gender   cauc hispanic other
  male     48        2    10
  female   33        2     4

, , sector = construction

        ethnicity
gender   cauc hispanic other
  male     19        0     3
  female    2        0     0

, , sector = other

        ethnicity
gender   cauc hispanic other
  male    169       12    26
  female  169       11    24

> ta<-xtabs(~gender+ethnicity+sector)
> chisq.test(ta)

	Chi-squared test for given probabilities

data:  ta
X-squared = 1573.753, df = 17, p-value < 2.2e-16
```