Vall d'Hebron
Institut de Recerca

Descriptive Statistics: Summaries and Graphs

Curs d'Estadística Bàsica per a la Recerca Biomèdica

UEB – VHIR
Santiago Pérez-Hoyos and Jordi Cortés
santi.perezhoyos@vhir.org

Vall d'Hebron Hospital · Generalitat de Catalunya Departament de Salut · Institut Català de la Salut · UAB Universitat Autònoma de Barcelona · Generalitat de Catalunya Departament d'Economia i Coneixement · Institució CERCA Centres de Recerca de Catalunya · ISC Instituto de Salud Carlos III · GRBIO

AIMS

- Make an approach to key concepts of **Statistics** and in particular to **Biostatistics**.
- Explain the different types of **analysis**, **variables** and other relevant **concepts**.
- Learn how to make a **statistical summary** of some descriptive data.
- Learn how to implement a descriptive statistics analysis with **R and R-Commander**.

Index

Index

GENERAL CONCEPTS

- **Population**: It represents the largest group of individuals who want to study and generally usually inaccessible.
- **Sample:** Subset of the population in which measurements are done. This sample should be representative of the original population (any individual has equal opportunity to be selected).
- **Variable**: Feature measurable and observable that represents a concept of study
- **Measure:** Procedure for assign quantitative or qualitative values to the characteristics of objects, people or events. If these procedures are not well measured the validity of the results is not guaranteed.

STEPS IN A STATISTICAL STUDY ANALSYIS

1. **Make hypothesis about a population**
2. **Decide which data collect (Experimental design)**
   - Which individuals will be part of the study (samples)
   - Which data must be collected for each individual (variables)
3. **Collect Data**
4. **Describe (summarize) collected data**
   - Summary measures and graphs
   - Point estimations and confidence intervals
5. **Establish relations between two variables**
   - Set up Statistical Hypothesis test
   - Check application conditions
   - Calculate intensity relationship measures
6. **Multivariable analysis . Modelling**
   - Consider effects of several variables on an outcome
   - Regression models
   - More complex models

Index

TYPES OF VARIABLES

QUALITATIVE

NOMINAL

Measure qualities of an individual

Examples: Sex, Treatment, Disease

ORDINAL

Measure qualities but they are ordered

Examples: Educational level, Stage, Severity

QUANTITATIVE

DISCRETE

Take only a finite possible values

Examples: Nº of admissions, Nº of programmed visits

CONTINOUS

Can take an infinite number of values. Between two measures always can be another

Example: Stay time, Age, Cholesterol level

Variable classification in a Study

- **Response, dependent or outcome variable**
  - One that answer the research question
- **Explain, independent or exposure variables**
  - They are those that are related to the causes of the events we want to study
- **Confounding or effect modifier variables**
  - Are those that can affect the relation between exposure and outcome variables
- **Universal variables**
  - Are those that can be exposures or confounders that always have to be considered. For example: sex, age, residence location, ethnic, etc.

Descriptive analysis

- Data have to be **organized** to be useful (frequency or contingency tables)
- **Graph** data before calculating summary measures
- This actions can help to:
  - Select the **best summary** measure
  - **Transform** variables
  - Detect **outliers**

Index

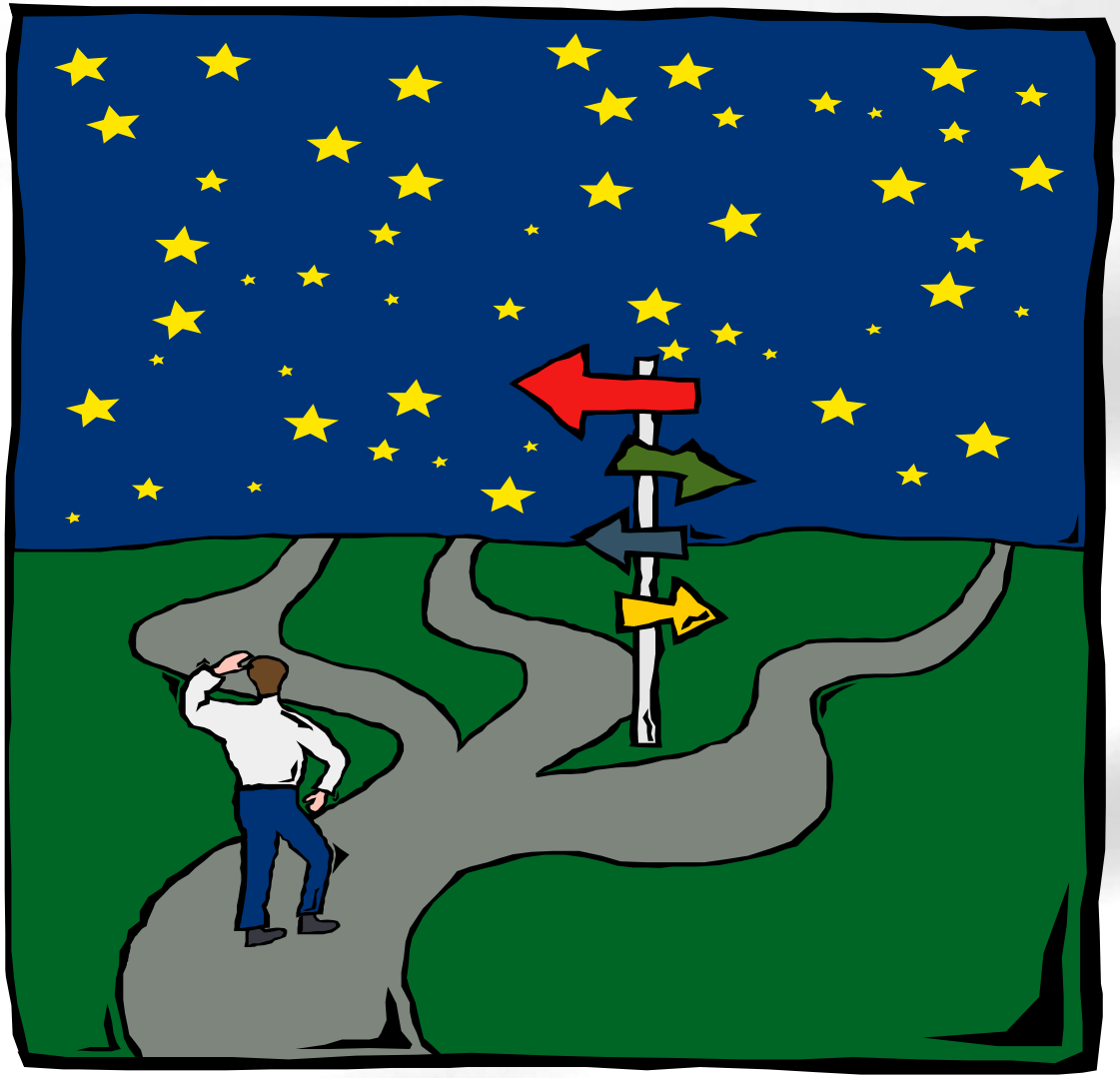
Quantitative Variables

- We have a new variable (i.e. a biomarker and we want to summarize information)
  - Around which values is the variable ?
  - Values vary greatly between different individuals
  - Data are grouped or not

Summary Measures

- Location
  - Mean
  - Median
  - Mode
- Dispersion
  - Range (Maximum-Minimum)
  - Variance
  - Standard Deviation
  - Variation Coefficient
  - Percentile
  - Interquartile range (IQR) or Interquartile interval
- Shape
  - Asymmetry
  - Kurtosis

Location measures



Mean

Median

Mode

Mean

- Useful to locate data .
- Is the **sum** of observed values **over sample size**
- Can be altered by extreme values

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Example Stay days

3, 4, 6, 9, 12 **Mean=6,8**

3, 4, 6, 9, 20 **Mean=8,4**

Mediana

- Is the point that divied in **two parts** the observations
- Observations are ordered from lowest to highest and median is the **central point**
- It is not altered by extreme observations

Example Stay days

3, 4, 6, 9, 12 **Median=6**

3, 4, 6, 9, 20 **Median=6**

Exemple en R-Commander

Exemple de Clase

Your Name

```
> x1<-c(3,4,6,9,12)
```

```
> x2<-c(3,4,6,9,20)
```

```
> dades<-data.frame(x1,x2)
```

```
> mean(x1)
```

```
[1] 6.8
```

```
> mean(x2)
```

```
[1] 8.4
```

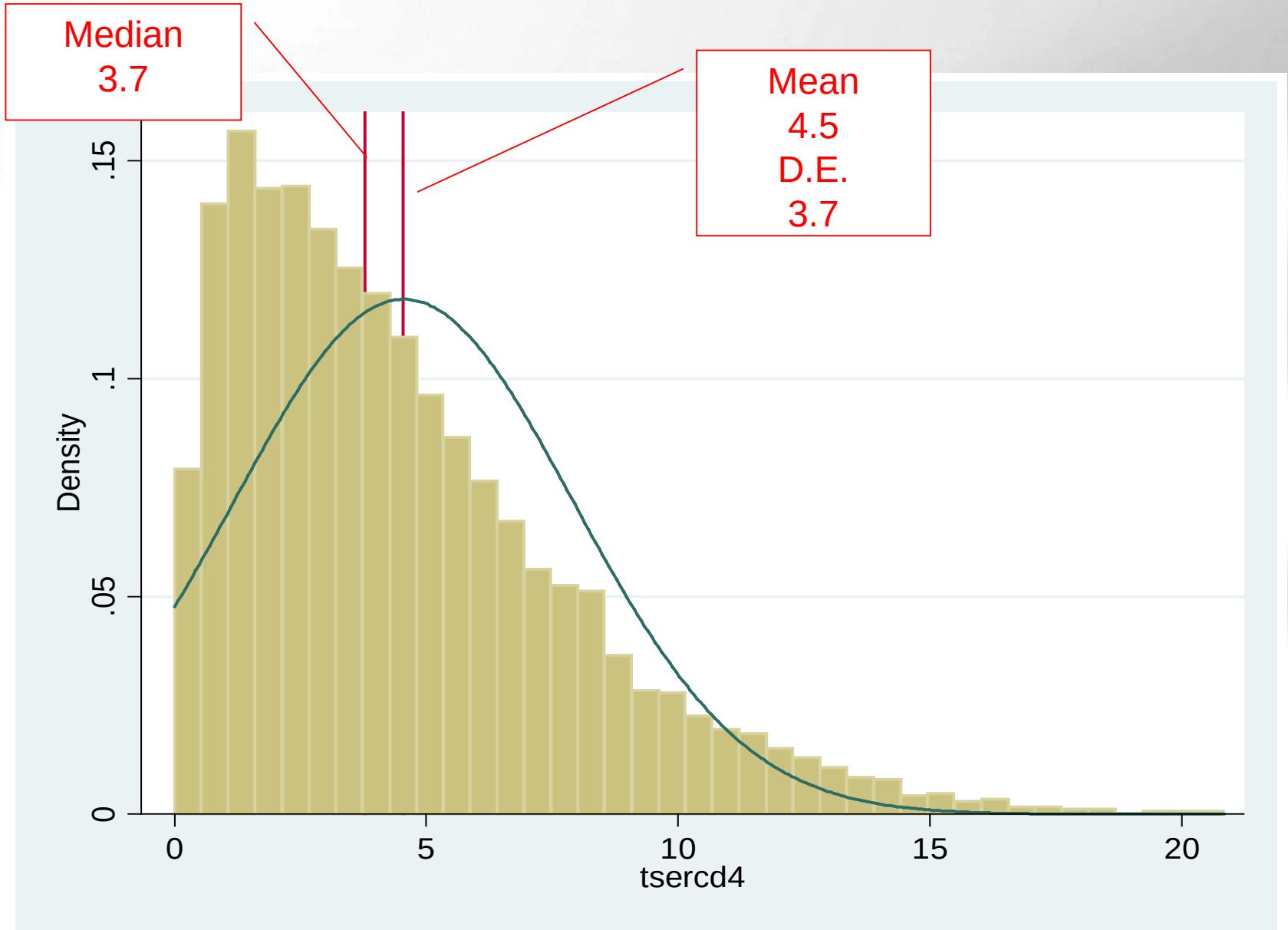
```
> median(x1)
```

```
[1] 6
```

```
> median(x2)
```

```
[1] 6
```

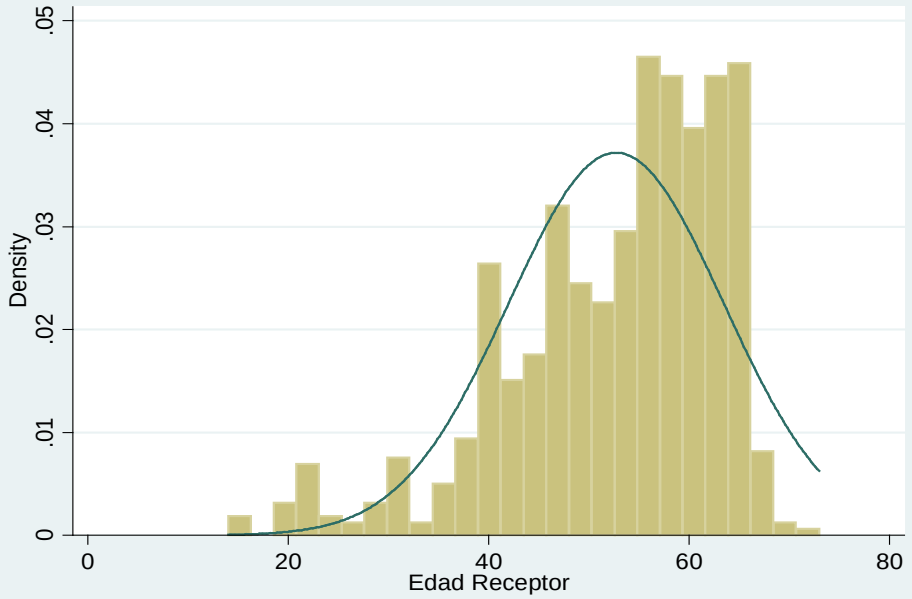
Mean or Median

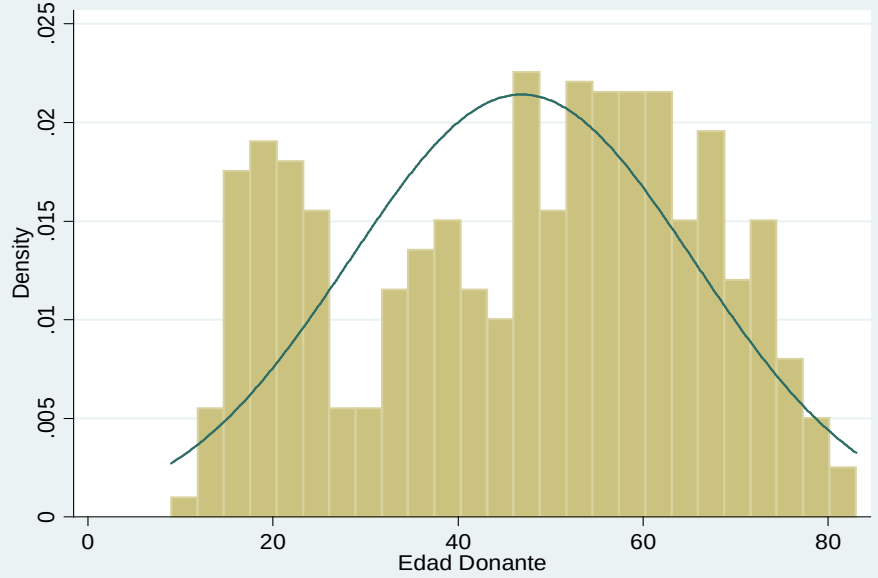


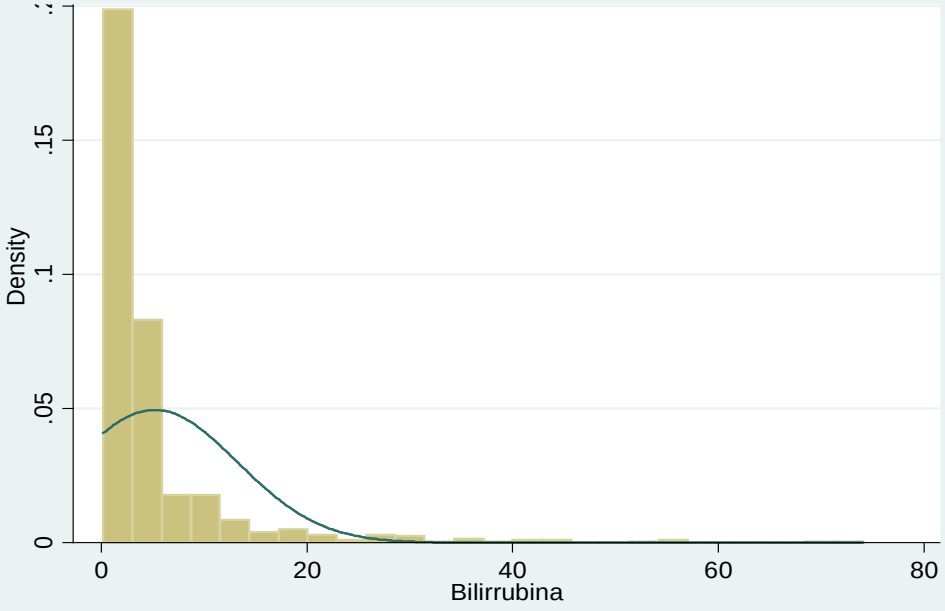
Vall d'Hebron
Institut de Recerca

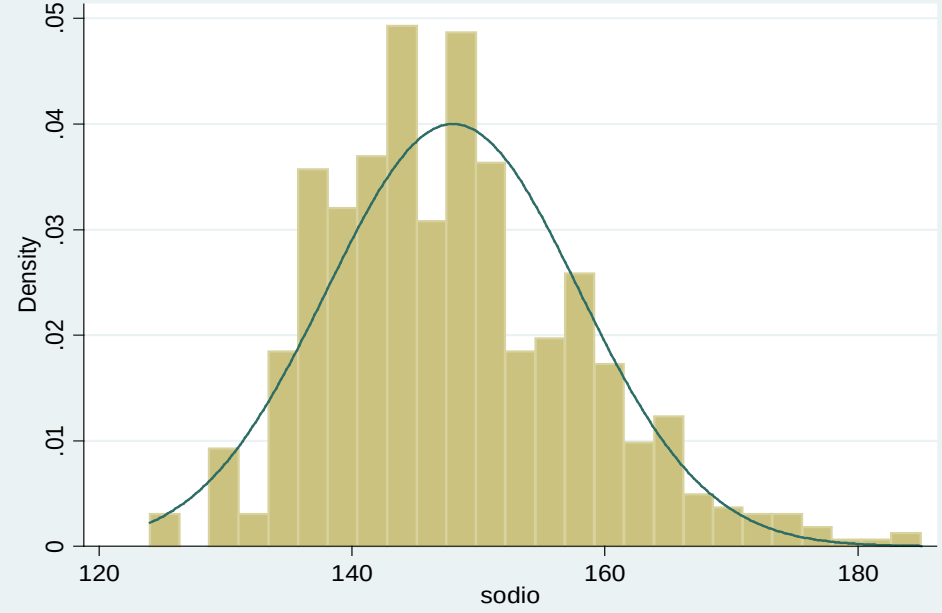
UNITAT
D'ESTADÍSTICA I
BIOINFORMÀTICA

Transplant study (real data)









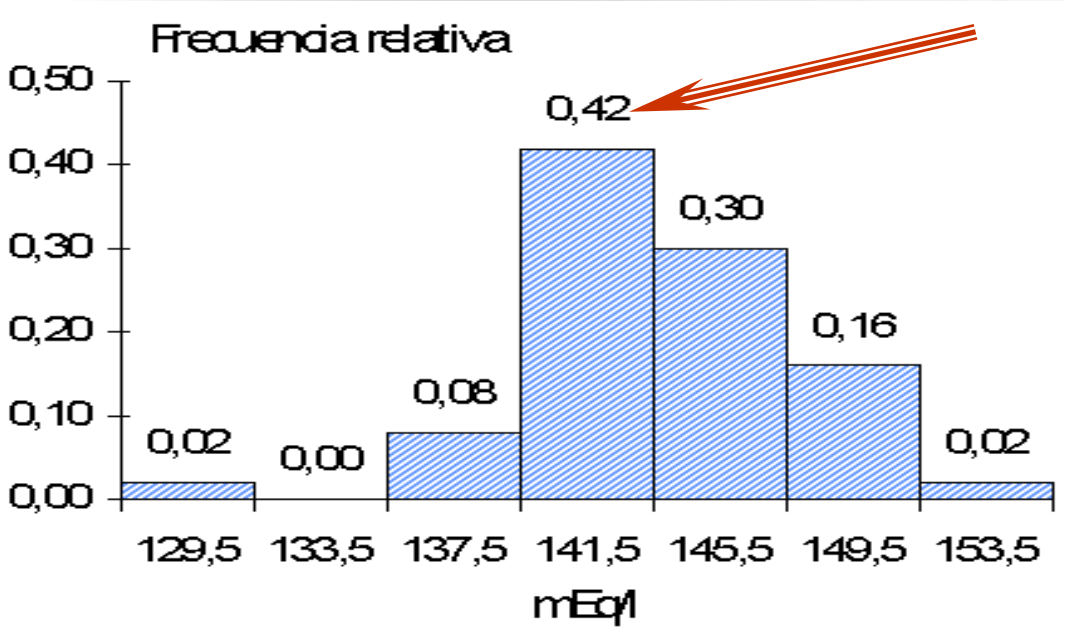
Geometric mean

- Sometimes data are transformed. For example logarithmic scale
- Mean is recalculated in transformed scale and exponentiated to come back to natural scale
- The calculated value is the **geometric mean**

	Days	Ln(Days)	Days	Ln(Days)
	3	1,10	3	1,10
	4	1,39	4	1,39
	6	1,79	6	1,79
	9	2,20	9	2,20
	12	2,48	20	3,00
Mean	6,8	1,79	8,4	1,89
Geometric Mean		6,00		6,65

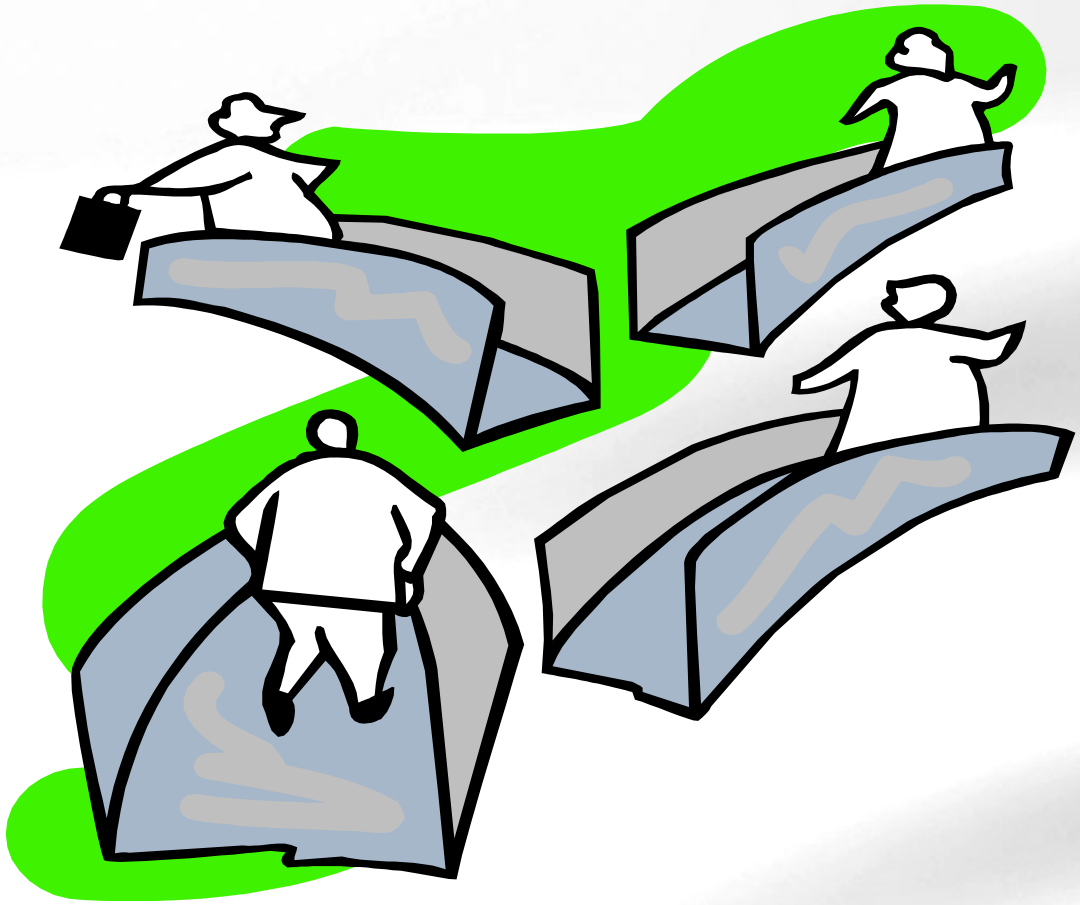
Mode

- The most frequent value
- May be not unique
- In a quantitative variable is the maximum values of an histogram



Determinaciones de sodio

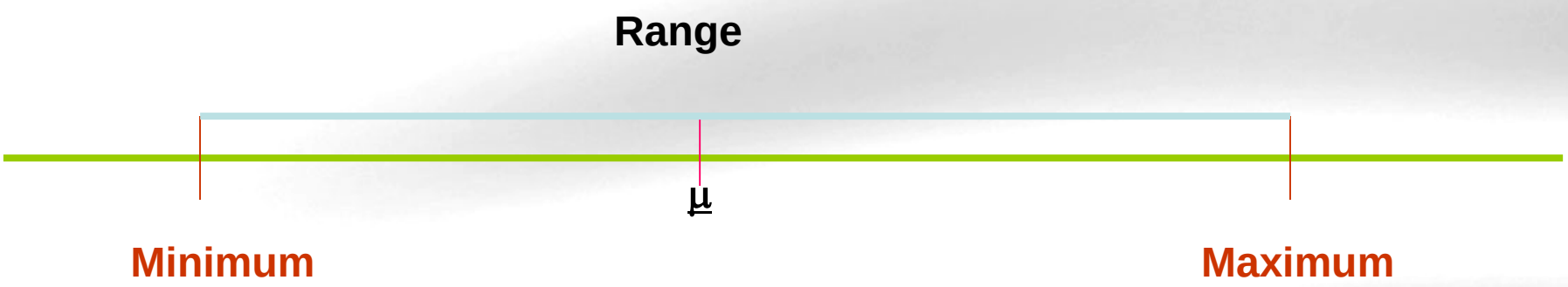
Dispersion or variability measures



- Range (Maximum-Minimum)
- Variance
- Standard Deviation
- Variation Coefficient
- Percentile
- Interquartile Range (IQR)

Range

- Simplest measure of dispersion
- Is the difference between maximum and minimum value of the observations



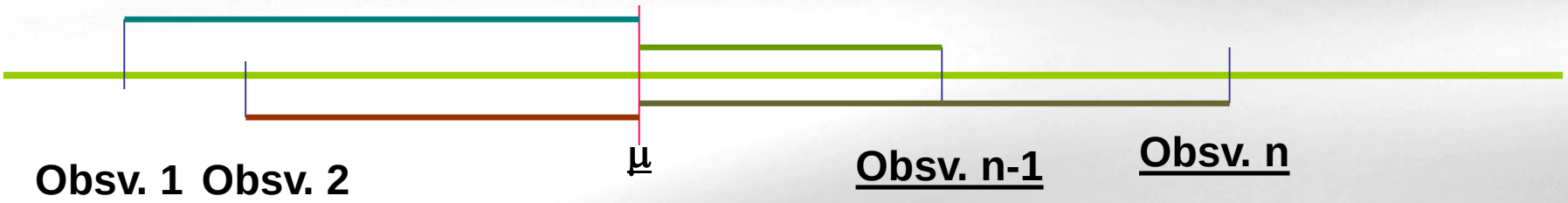
Example Stay days

3, 4, 6, 9, 12 **Range=12-3=9**

3, 4, 6, 9, 20 **Range=20-3=17**

Variance

- Mean difference of observations from mean in squared scale



	d	d-mean	(d-mean)^2	d	d-mean	(d-mean)^2
	3	-3,8	14,44	3	-5,4	29,16
	4	-2,8	7,84	4	-4,4	19,36
	6	-0,8	0,64	6	-2,4	5,76
	9	2,2	4,84	9	0,6	0,36
	12	5,2	27,04	20	11,6	134,56
Sum		0	54,8		0	189,2
Sum/5	6,8	0	10,96	8,4	0	37,84

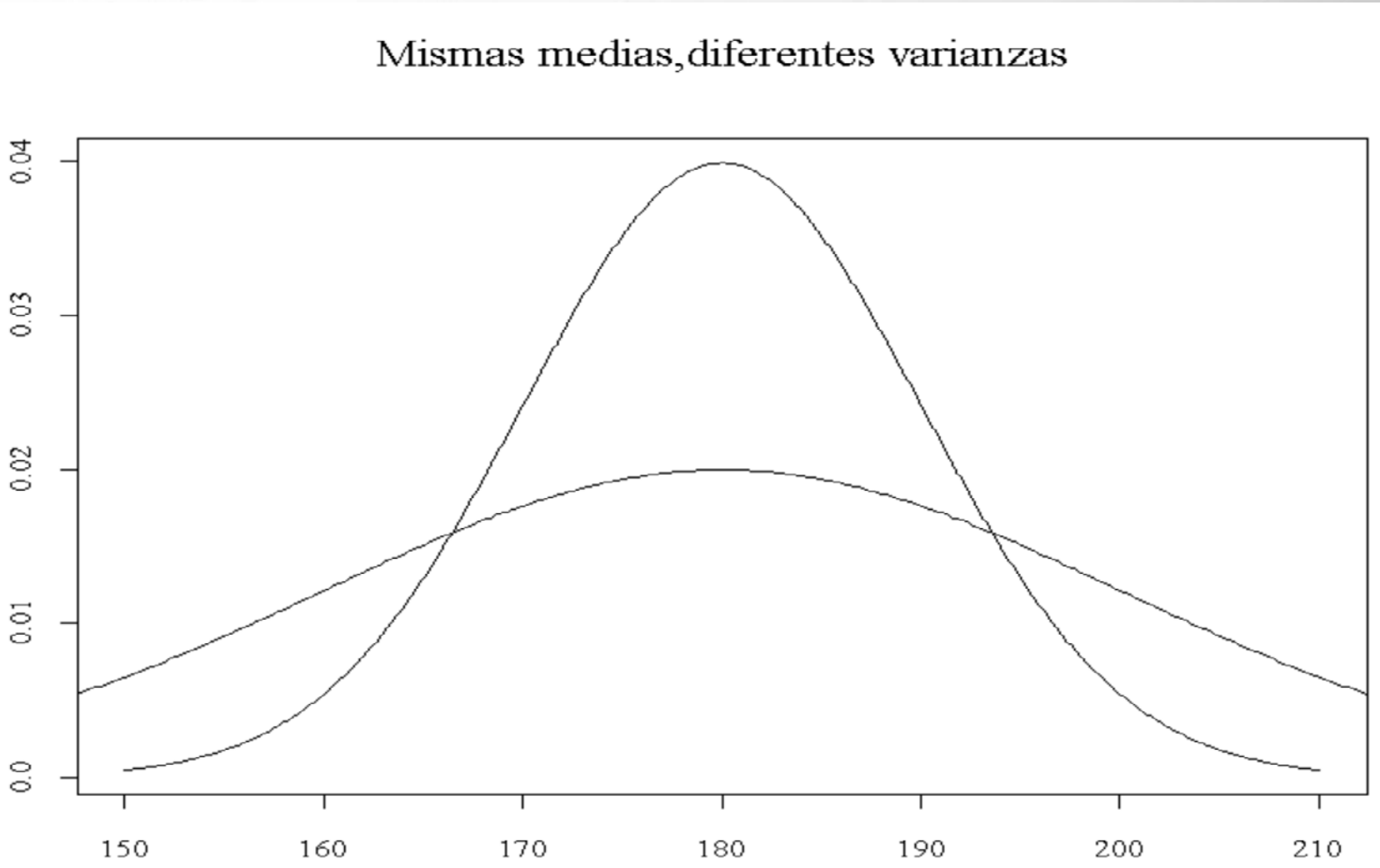
Standard Deviation

- Squared root of the variance
- It is measured in the same units than the variable

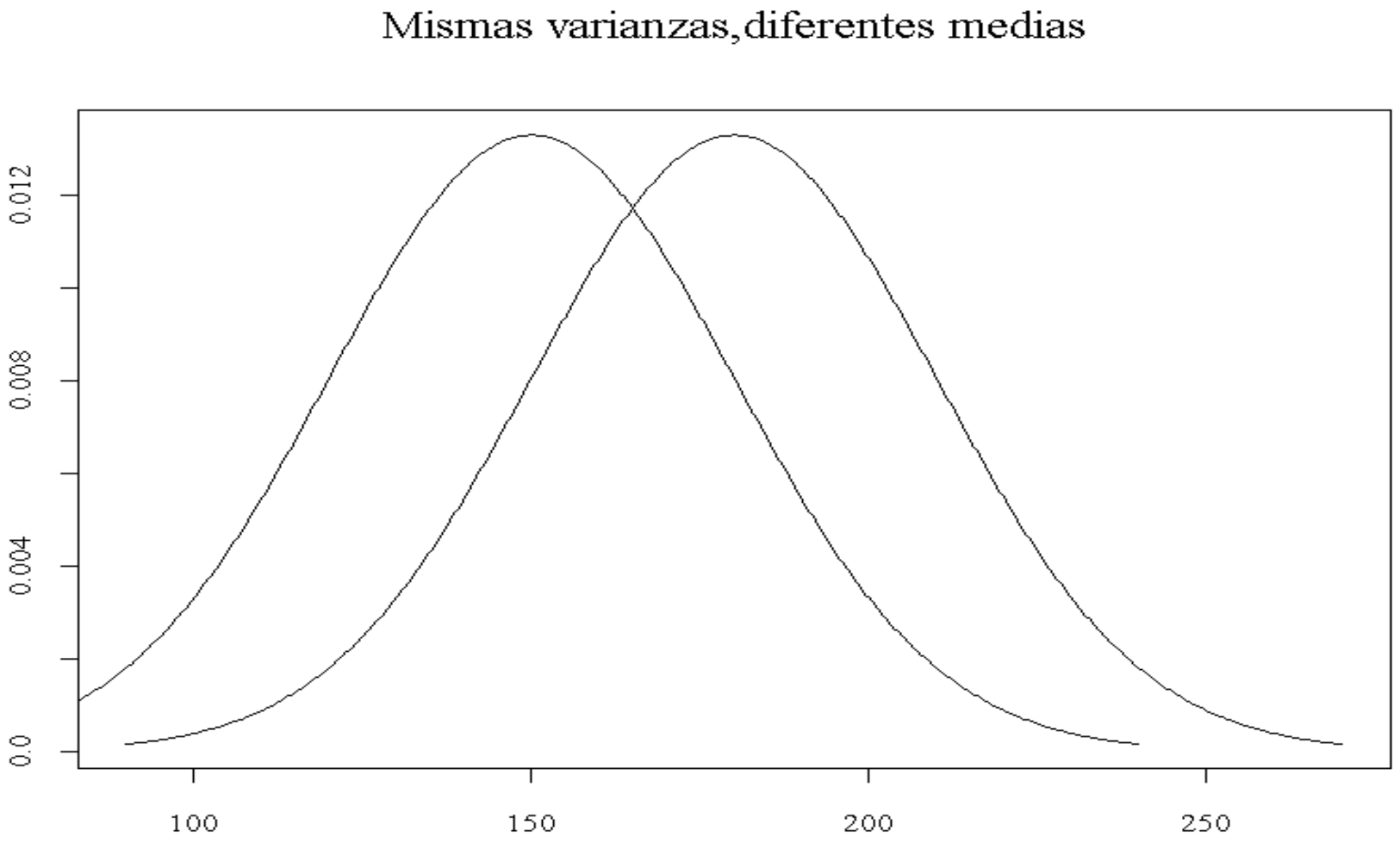
Example Stay days

3, 4, 6, 9, 12	**Variance =10.96**
	Std. Dev.=3.31
3, 4, 6, 9, 20	**Variance= 37.84**
	Std. Dev.= 6.15

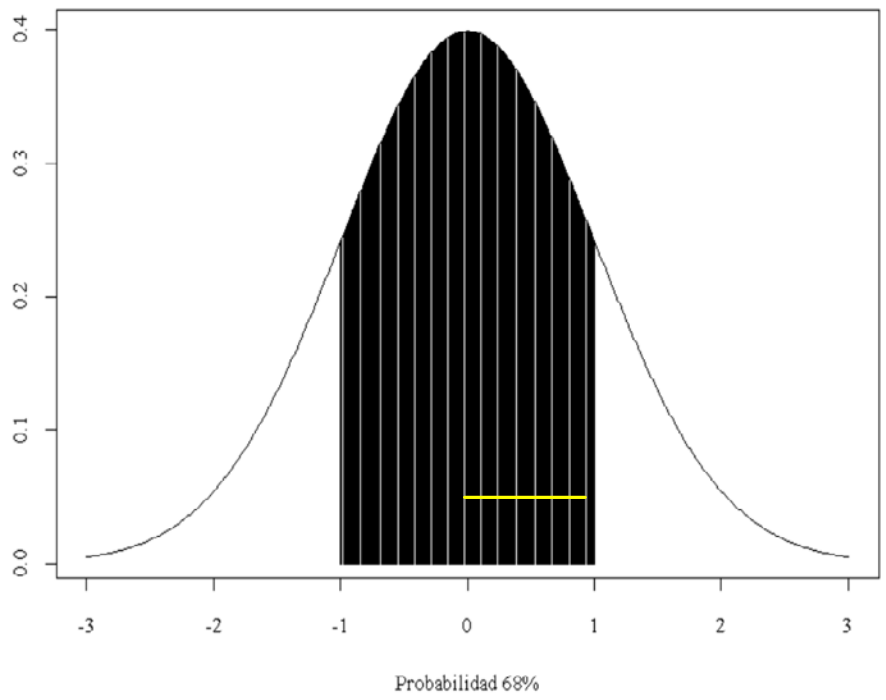
Same mean , different variances

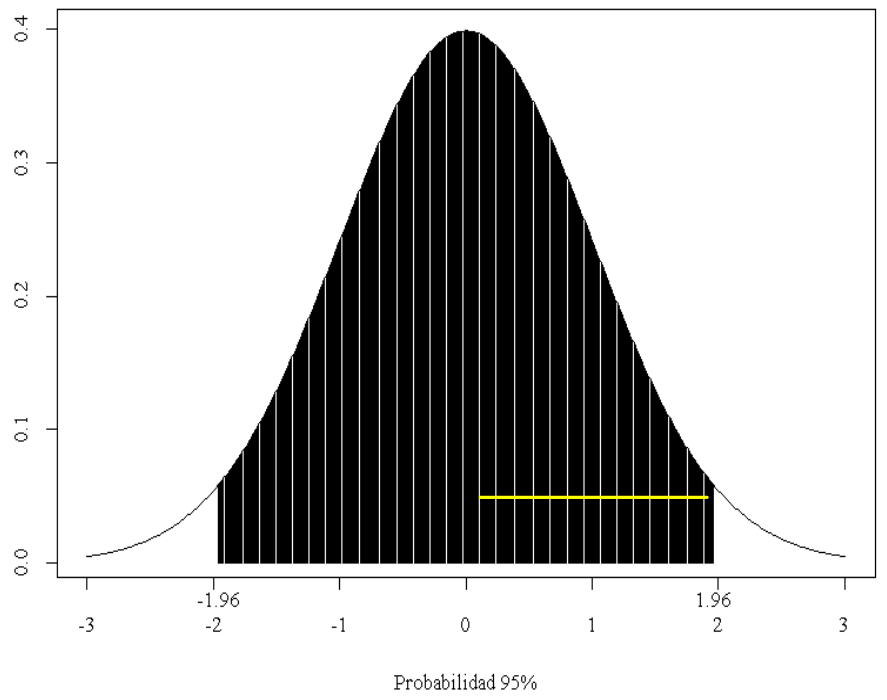


Same variances, different means



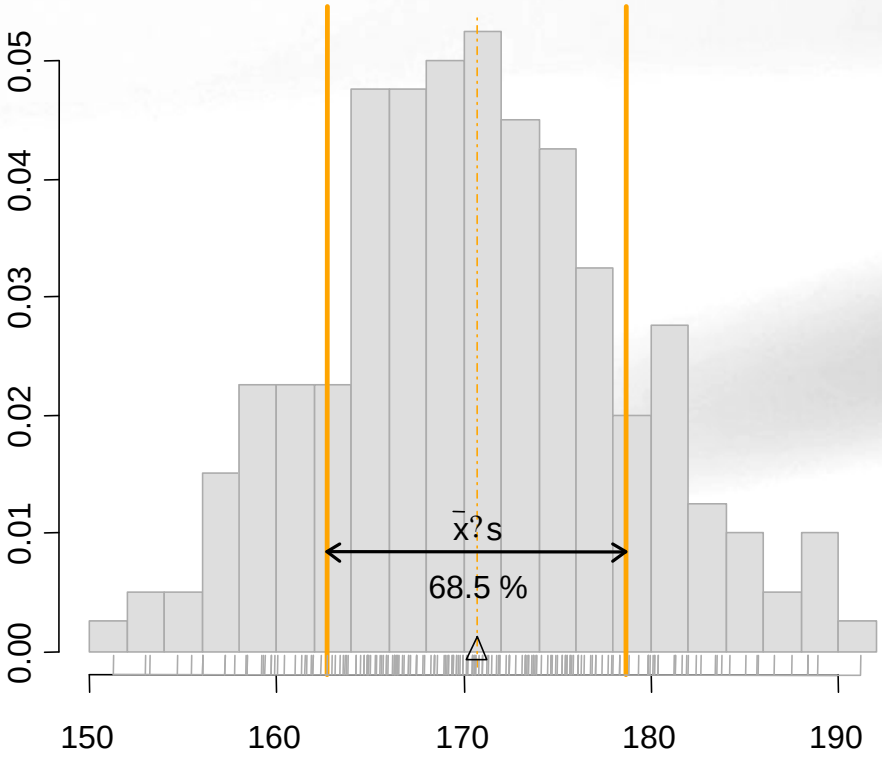
Normal distribution

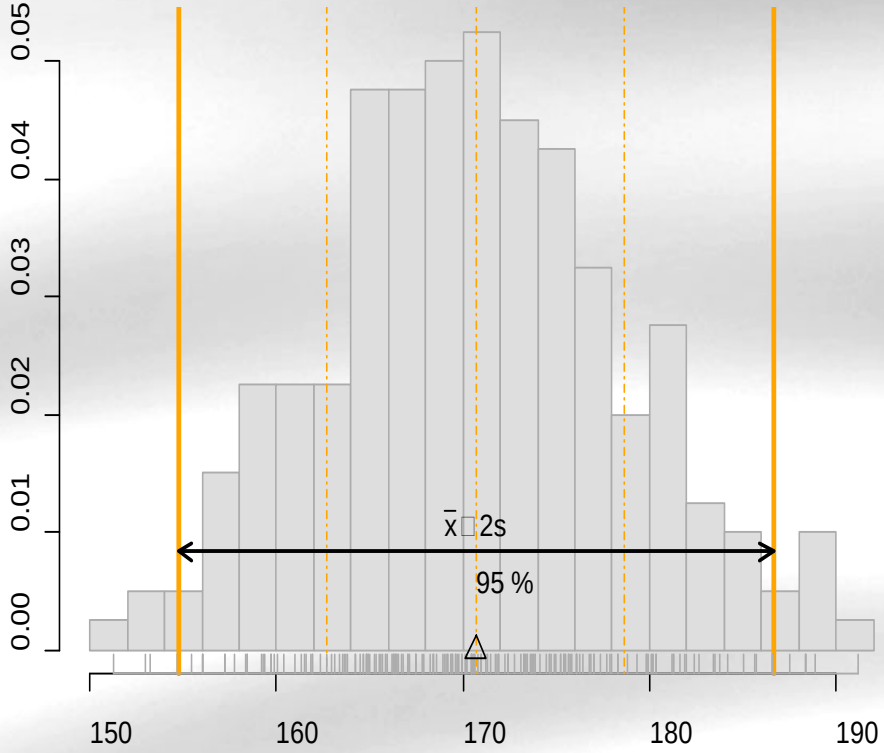




- ✓ Between μ (mean) ± 1 SD, there are about 68% of the probability.
- ✓ Between μ (mean) ± 2 SD is about 95% of probability.

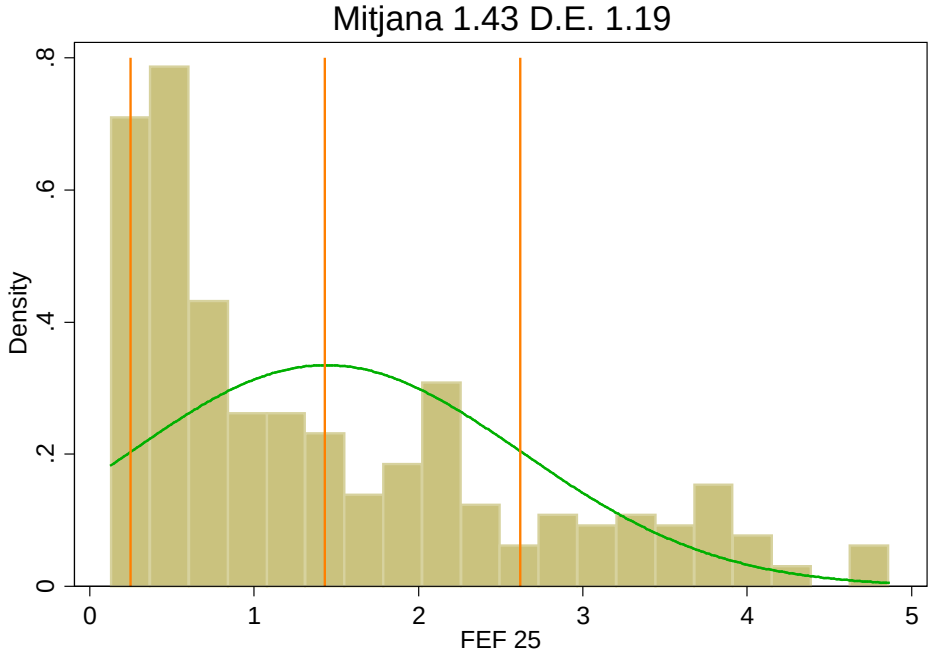
Symmetric distribution of data

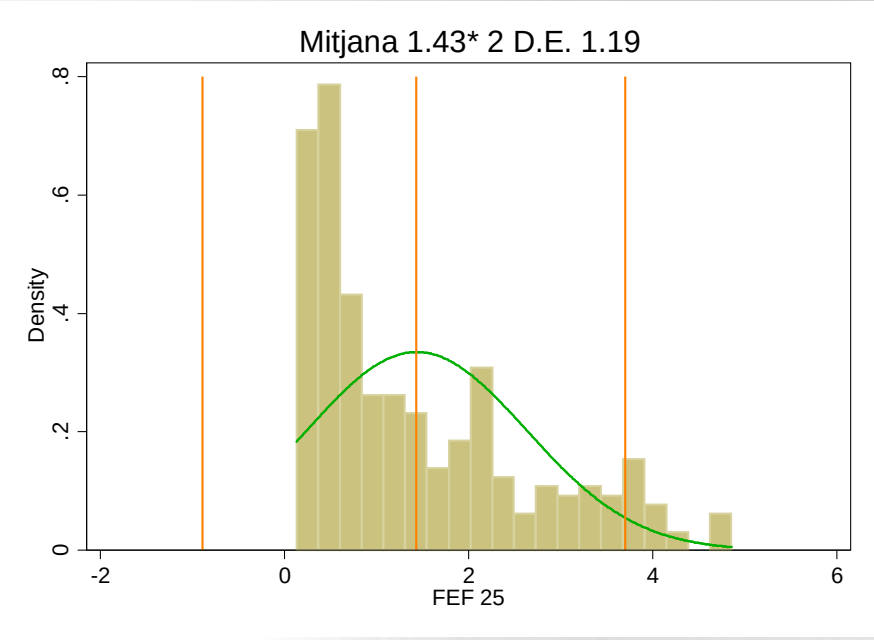




- ✓ Between $\bar{x}$ (mean) ± 1 SD, there are about 68% of the observations.
- ✓ Between $\bar{x}$ (mean) ± 2 SD is about 95% of observations

Asymmetric distribution of data





- ✓ Between $\bar{x}$ (mean) ± 1 SD, there percentage of observations is not 68%
- ✓ Between $\bar{x}$ (mean) ± 2 SD, there percentage of observations is not 95%

Coefficient of Variation (CV)

- It is the ratio between standard deviation and mean
- Allows to compare the variability of variables measured in different scales

Example Stay Days

3, 4, 6, 9, 12	**Std. Dev.= 3.31** **Mean = 6.8** **Variation Coef.= 0.49**
3, 4, 6, 9, 20	**Std. Dev. = 6.15** **Mean = 8.4** **Variation Coef.= 0.73**

Example Variation Coefficient

Weight (Kg) of 5 patients → (70,60,56,83,79)

$\bar{x}$ = Kg

s = Kg

Blood Pression (mmHg) → 5 patients (150,170,135,180,19)

$\bar{x}$ = mmHg

s = mmHg

Which variable have more variation? Weight or Pression

Calculate CV

CV_{WGT} =

CV_{BP} = %

R Commander

File Edit Data Statistics Graphs Models Distributions Too

Data set: dades | Edit data set | View data set

R Script | R Markdown

```
x1<-c(70,60,56,83,79)
x2<-c(150,170,135,180,195)
dades<-data.frame(x1,x2)
mean(x1)
sd(x1)
mean(x2)
sd(x2)
(sd(x1)/mean(x1))*100
(sd(x2)/mean(x2)*100
```

Example Variation Coefficient

Weight (Kg) of 5 patients → (70,60,56,83,79)

- X = 69.6 Kg
- s = 11.67

Blood Pression (mmHg) 5 patients → (150,170,135,180,195)

- X = 166 mmHg
- s = 23.82

Which variable have more variation? **Weight** or Pression

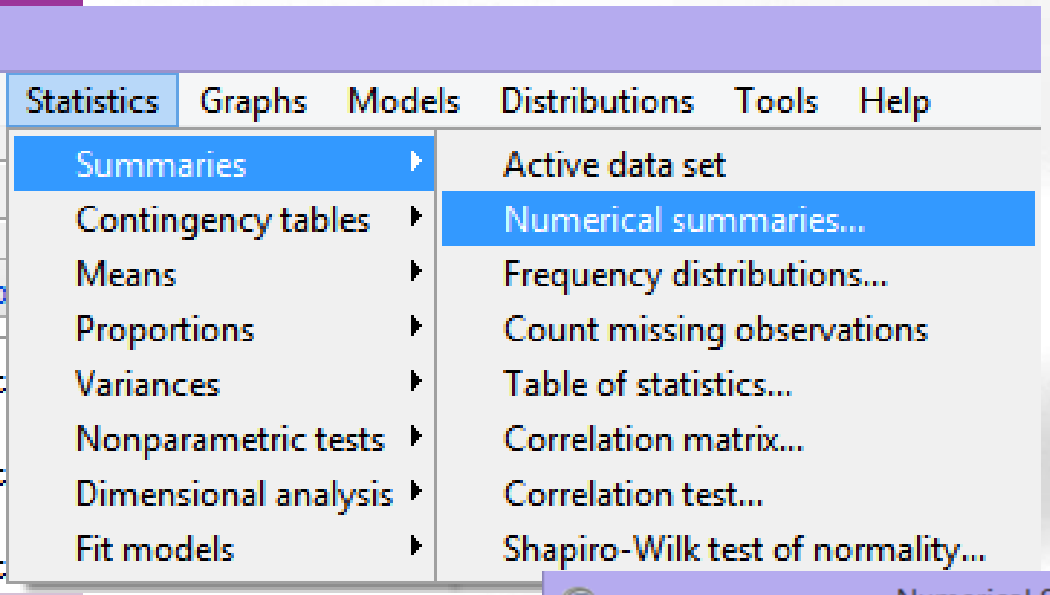
Calculate CV

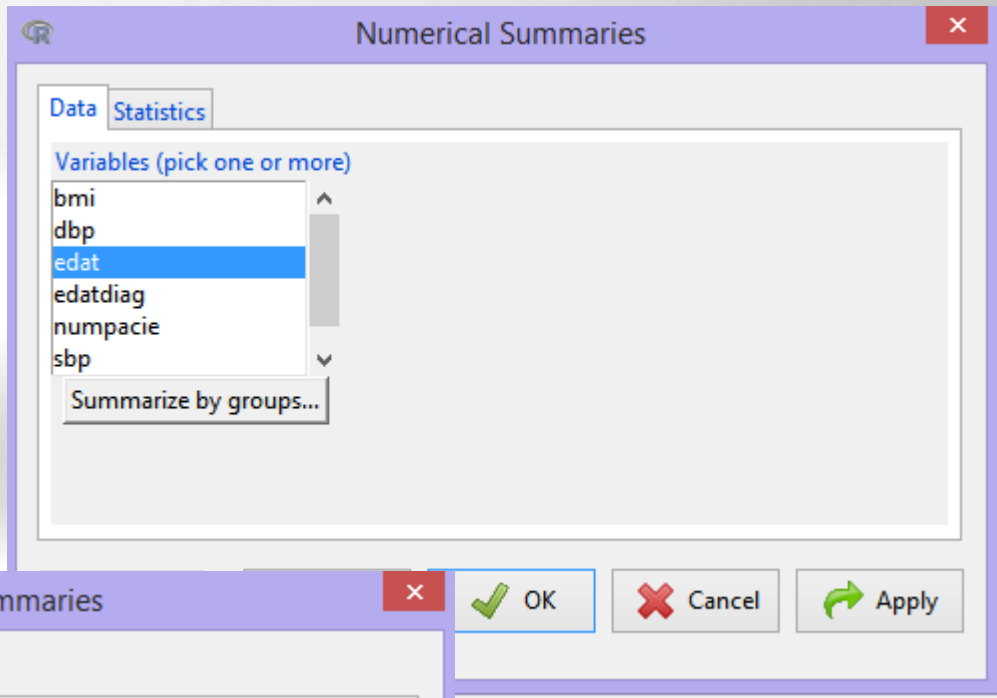
- CV_{PES} = 16.7%
- CV_{TA} = 14.3%

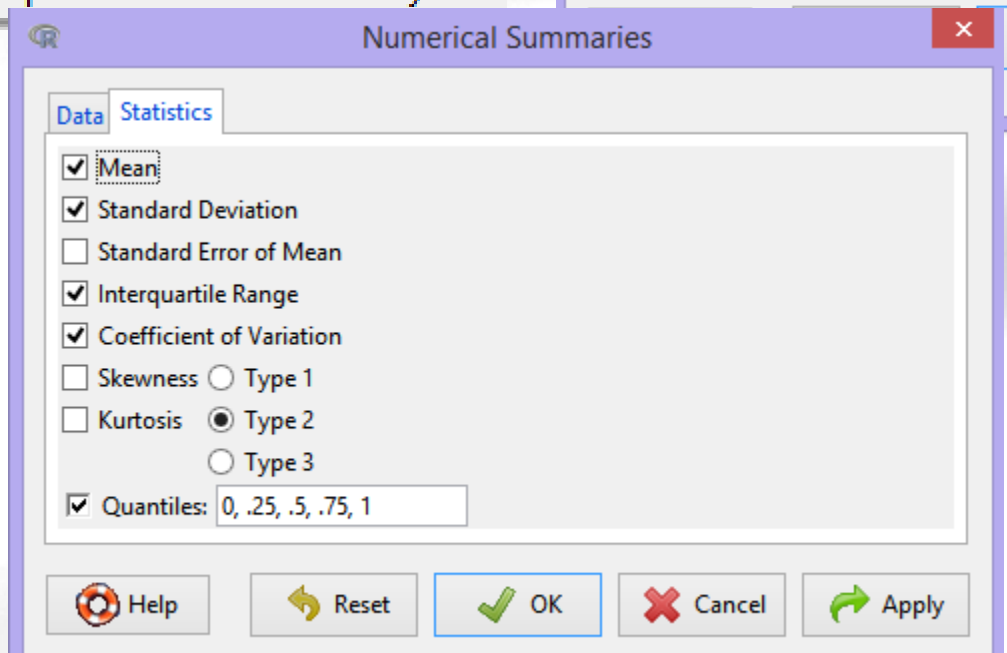
Percentiles

- Observations are ranked from minimum to maximum and the point that leaves below p% of observations is selected
- There are some special percentiles
  - Deciles are percentiles 10, 20, 30, 40, 50, 60, 70, 80, 90
  - Quartiles are percentiles 25, 50, 75
  - Quintiles are percentiles 20,40, 60,80
  - They are not affected by extreme observations
  - Interquartile range is difference between 25 and 75 percentile

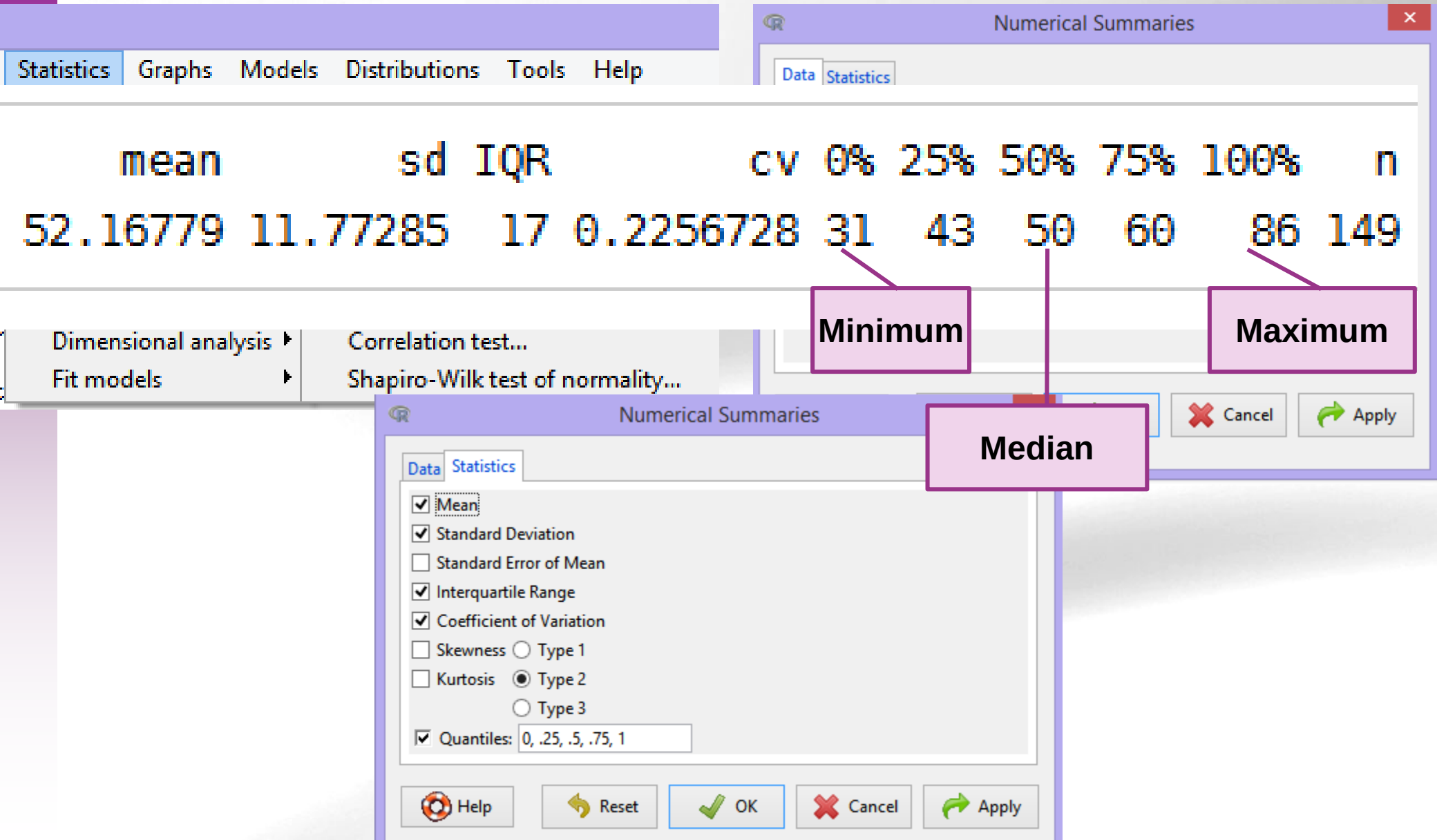
Summary meaures in R Commander







Summary measures in R Commander

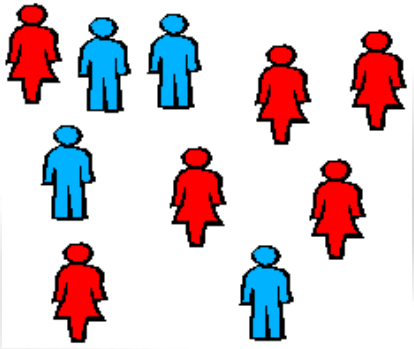


Syllabus

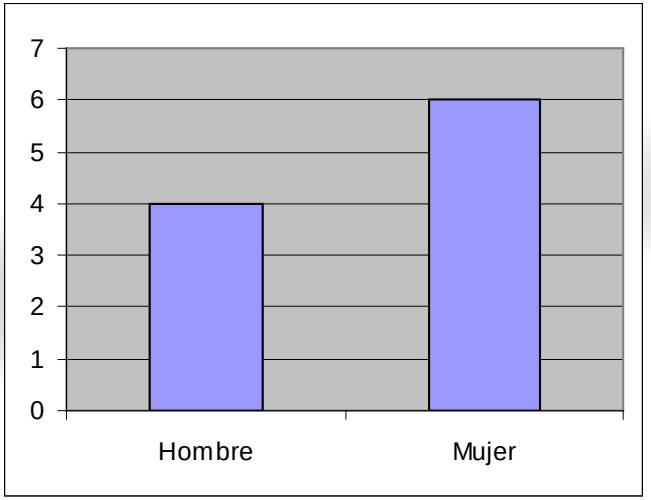
1. INTRODUCTION. ANALYSIS STRATEGY
2. VARIABLES CLASSIFICATION
3. SUMMARY MEASURES
    1. Measures of location/central tendency
    2. Measures of variability/dispersion
4. SUMMARY OF VARIABLES
    1. Contingency Tables
    2. Graphs
5. Examples & exercises

Summary of variables

Frequency tables and graphs are two equivalent ways to present information. Both expose in an ordered way the collected data.a.



Género	Frec.
Hombre	4
Mujer	6



Frequency table

NÚMERO DE HIJOS ENTRE MUJERES DE 20 Y 30 AÑOS

NÚMERO DE HIJOS	Frecuencia Absoluta (f_i)	Frecuencia Relativa (fr_i)	Frecuencia Acumulada (F_i)	Frecuencia Relativa Acumulada (Fr_i)
0	175	0'35	175	0'35
1	225	0'45	400	0'80
2	75	0'15	475	0'95
3 o más	25	0'05	500	1'00
TOTAL	500	1'00	500	1'00

Frequency table

Cate
go
ries

NÚMERO DE HIJOS ENTRE MUJERES DE 20 Y 30 AÑOS

NÚMERO DE HIJOS	Frecuencia Absoluta (f_i)	Frecuencia Relativa (fr_i)	Frecuencia Acumulada (F_i)	Frecuencia Relativa Acumulada (Fr_i)
0	175	0'35	175	0'35
1	225	0'45	400	0'80
2	75	0'15	475	0'95
3 o más	25	0'05	500	1'00
TOTAL	500	1'00	500	1'00

Taula de Freqüencia

Cate go ries

Nº of subjects by category

NÚ… …ENTRE MUJERES DE 20 Y 30 AÑOS

NÚMERO DE HIJOS	Frecuencia Absoluta (f_i)	Frecuencia Relativa (fr_i)	Frecuencia Acumulada (F_i)	Frecuencia Relativa Acumulada (Fr_i)
0	175	0'35	175	0'35
1	225	0'45	400	0'80
2	75	0'15	475	0'95
3 o más	25	0'05	500	1'00
TOTAL	500	1'00	500	1'00

Frequency table

Percentage of subjects
Freq /Total

Nº of subjects by category

Cate go ries

NÚ… ENTRE MUJERES DE 20 Y 30 AÑOS

NÚMERO DE HIJOS	Frecuencia Absoluta (f_i)	Frecuencia Relativa (fr_i)	Frecuencia Acumulada (F_i)	Frecuencia Relativa Acumulada (Fr_i)
0	175	0'35	175	0'35
1	225	0'45	400	0'80
2	75	0'15	475	0'95
3 o más	25	0'05	500	1'00
TOTAL	500	1'00	500	1'00

Frequency table

Percentage of subjects
Freq /Total

Nº of subjects by category

Cate go ries

NÚ… ENTRE MUJERES DE 20 Y 30 AÑOS

NÚMERO DE HIJOS	Frecuencia Absoluta (f_i)	Frecuencia Relativa (fr_i)	Frecuencia Acumulada (F_i)	Frecuencia Relativa Acumulada (Fr_i)
0	175	0'35	175	0'35
1	225	0'45	400	0'80
2	75	0'15	475	0'95
3 o más	25	0'05	500	1'00
TOTAL	500	1	500	1'00

Nº accumulated subects up to category
(Only ordinal or discrete variables)

Frequency table

Percentage of subjects
Freq /Total

Nº of subjects by category

Cate go ries

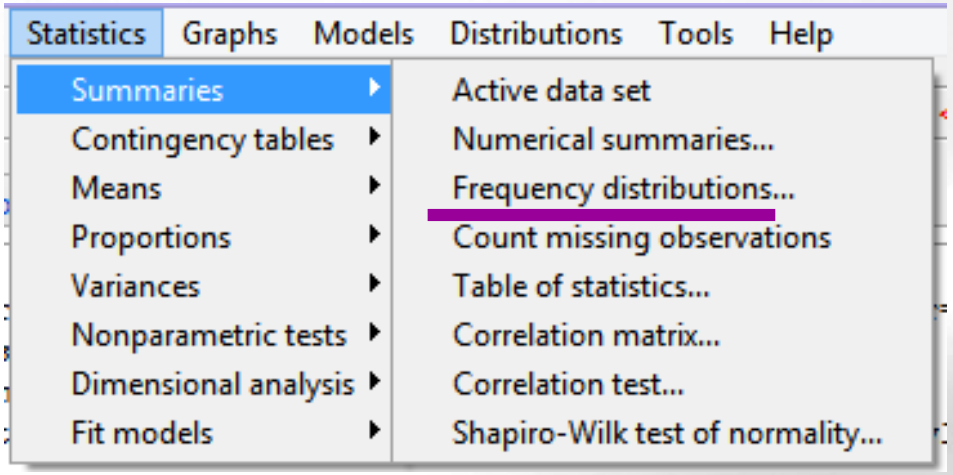
NÚ… ENTRE MUJERES DE 20 Y 30 AÑOS

NÚMERO DE HIJOS	Frecuencia Absoluta (f_i)	Frecuencia Relativa (fr_i)	Frecuencia Acumulada (F_i)	Frecuencia Relativa Acumulada (Fr_i)
0	175	0'35	175	0'35
1	225	0'45	400	0'80
2	75	0'15	475	0'95
3 o más	25	0'05	500	1
TOTAL	500	1	500	

Nº accumulated subects up to category
(Only ordinal or discrete variables)

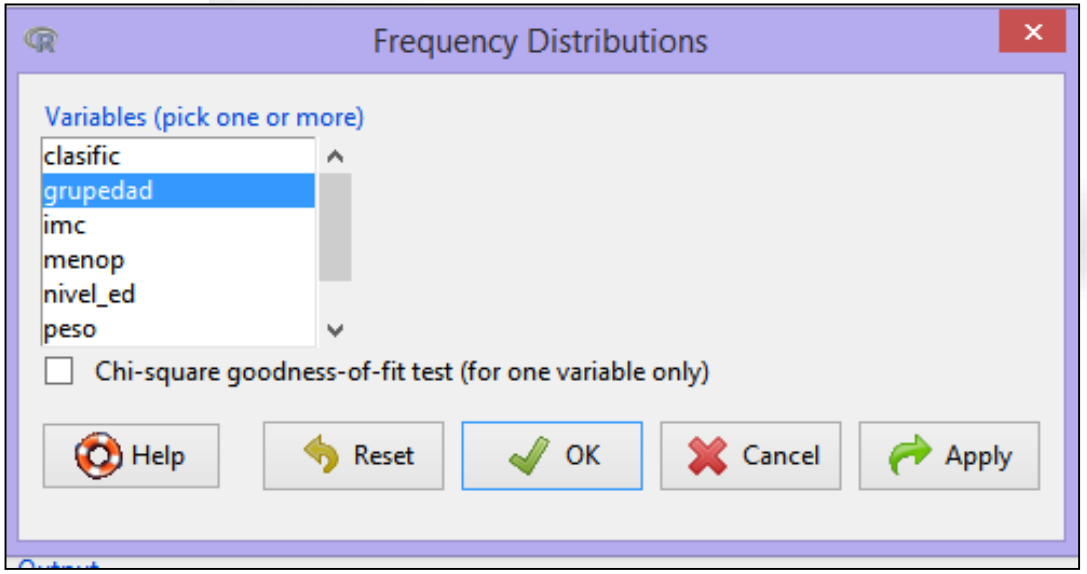
Accumulated Frequency up to category
Freq Abs/Total

Frequency tables in R Commander




```
counts:
grupedad
45 - 49 50 - 54 55 - 59 60 - 64 65 - 69
    378     233     176     129      84

percentages:
grupedad
45 - 49 50 - 54 55 - 59 60 - 64 65 - 69
   37.8    23.3    17.6    12.9     8.4
```



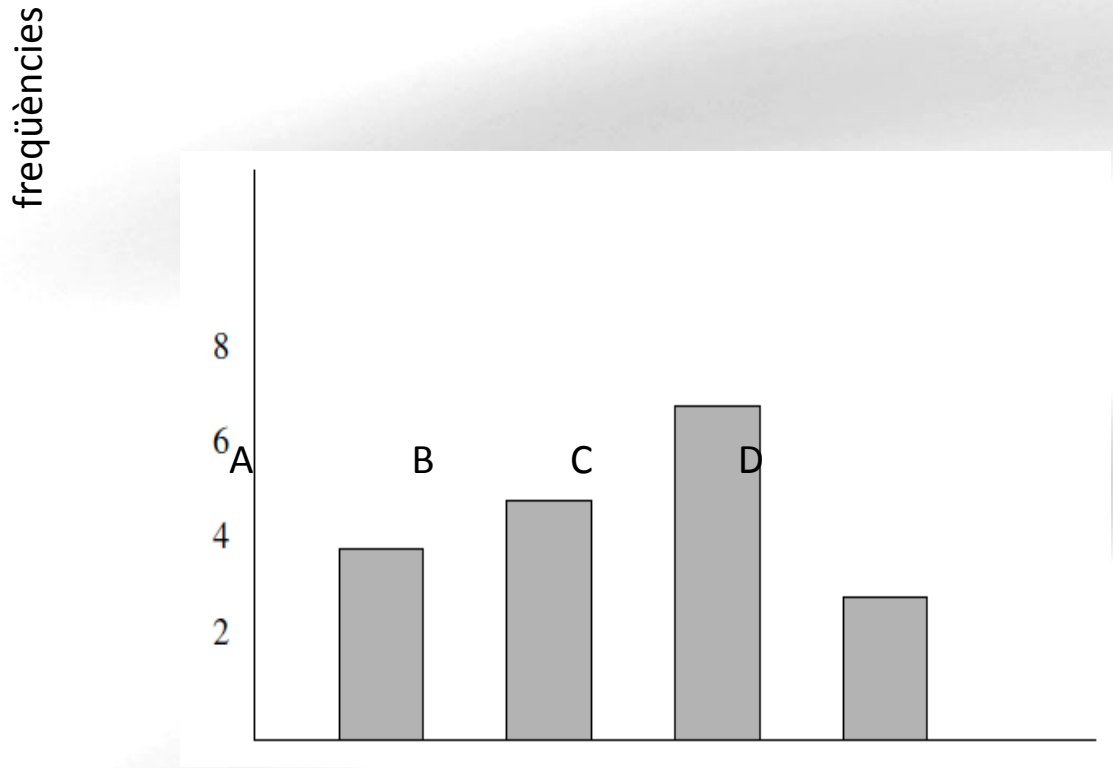
Frequency tables for quantitative variables

PESO

Marca de Clase	Intervalo de Clase	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Acumulada	Frecuencia Relativa Acumulada
42'5	x<45	1	0'002	1	0'002
47'5	45<=x<50	3	0'006	4	0'008
52'5	50<=x<55	12	0'024	16	0'032
57'5	55<=x<60	75	0'150	91	0'182
62'5	60<=x<65	103	0'206	194	0'388
67'5	65<=x<70	155	0'310	349	0'698
72'5	70<=x<75	101	0'202	450	0'900
77'5	75<=x<80	29	0'058	479	0'958
82'5	80<=x<85	11	0'022	490	0'980
87'5	85<=x<90	8	0'016	498	0'996
92'5	90<=x<95	2	0'004	500	1'000
		500	1'000	500	1'000

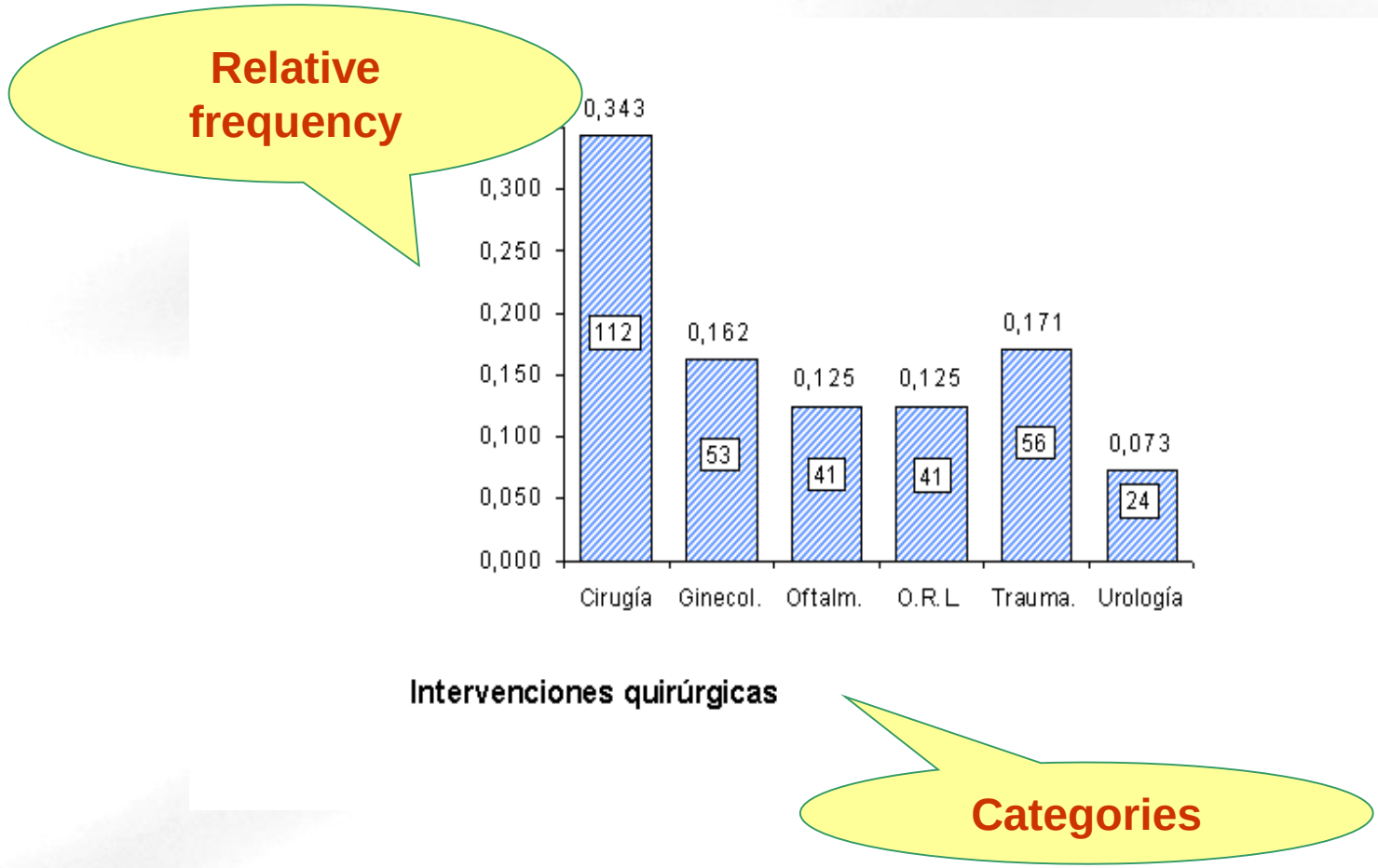
Bar Graph

Categorys are representened in X axis and frequencies in Y axis

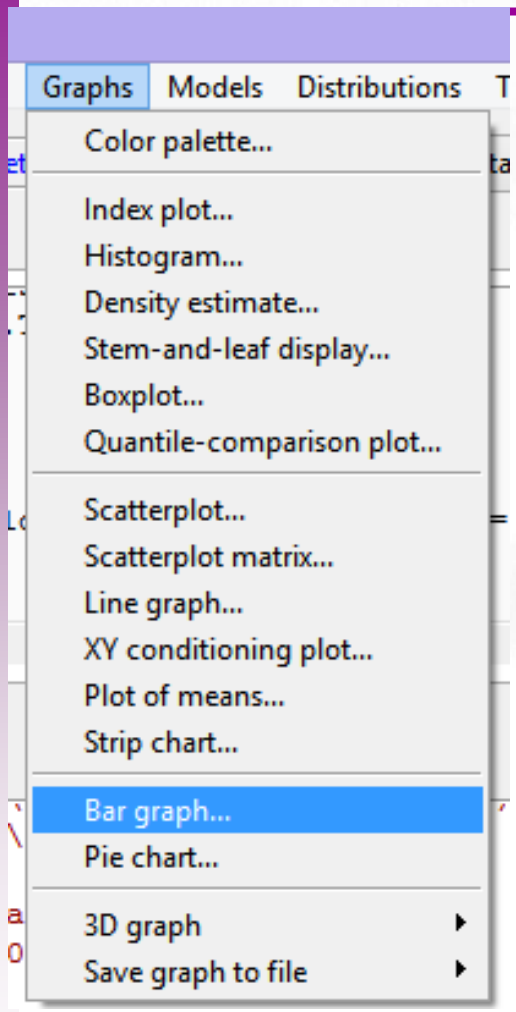


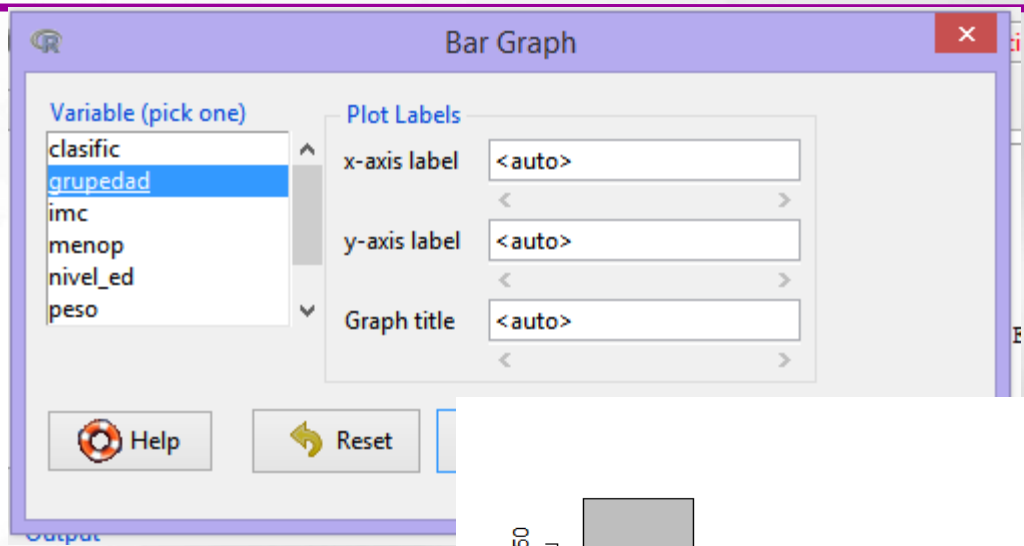
Bar Graph

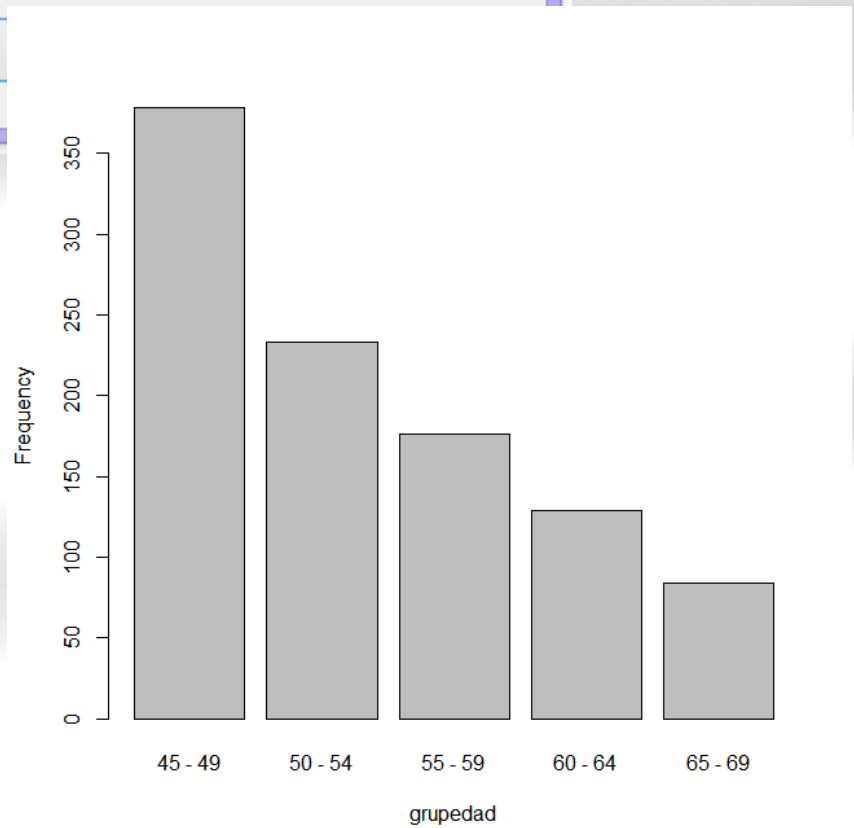
For comparing two population better use relative frequencies



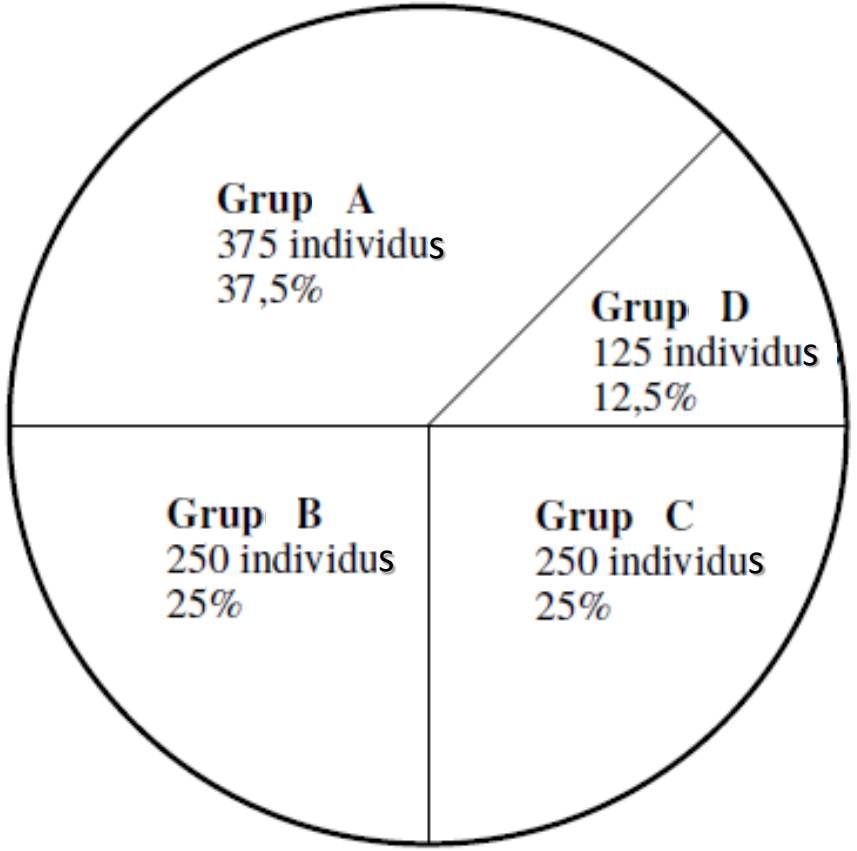
Graphs in R Commander







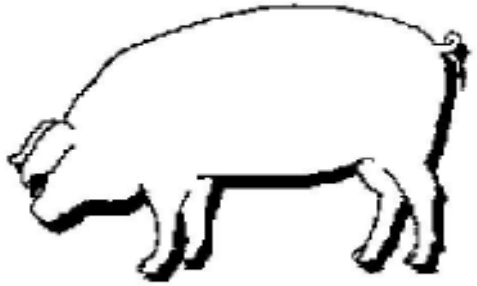
Pie Graph


Grup A
375 individus
37,5%
Grup D
125 individus
12,5%
Grup B
250 individus
25%
Grup C
250 individus
25%

Pictograms

Expressed with drawings alluding to the subject of study frequencies of the modalities of the variable. The scaling of each design should be such that the area of each of them should be proportional to the frequency category representing. Used by the media because they can be quickly understood by a general audience..

Ampolles de cervesa recollides



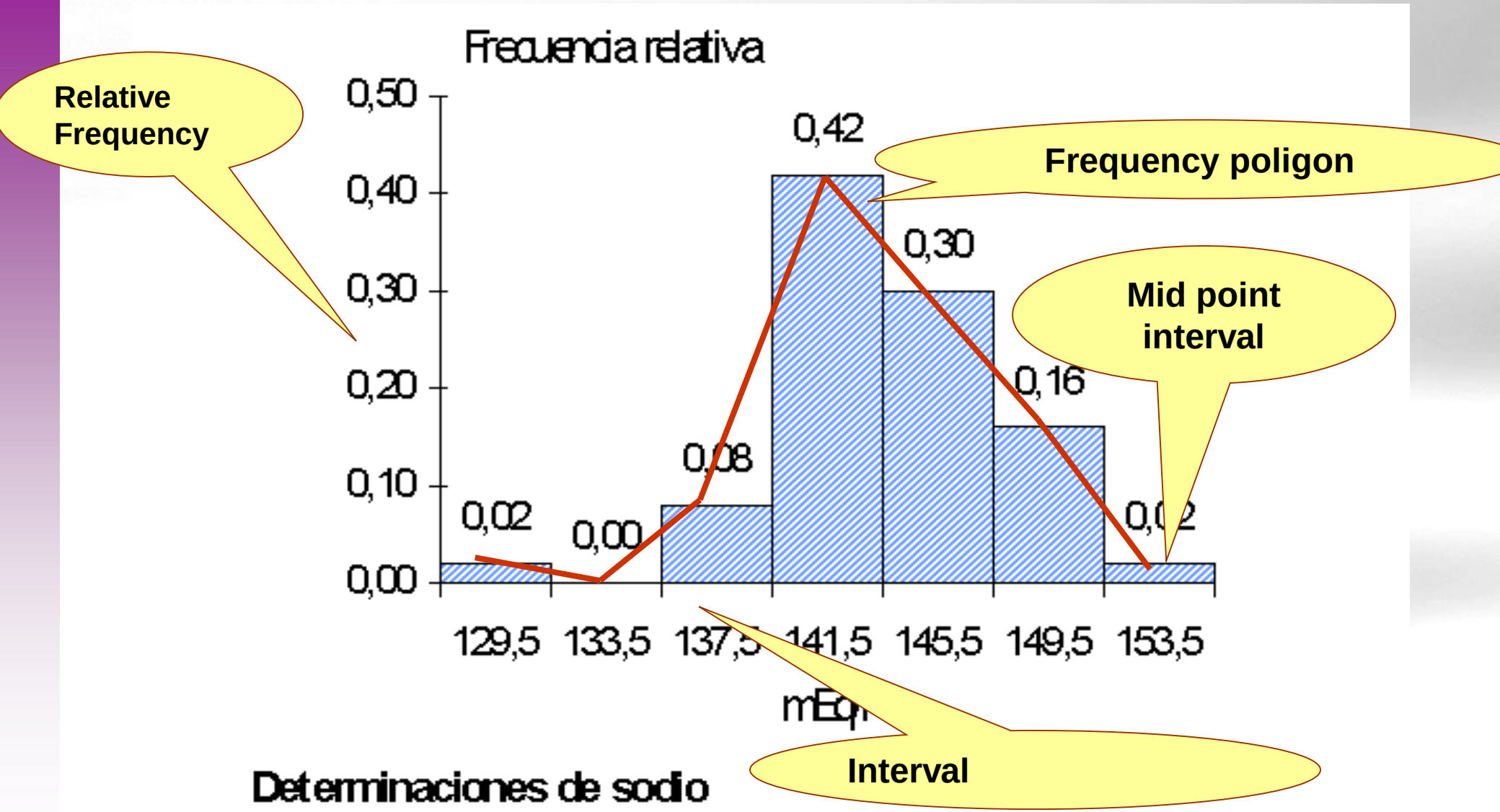
100 Kg

Ciutat A

50Kg

Ciutat B

Histogram


Relative Frequency
Frecuencia relativa
0,50
0,40
0,30
0,20
0,10
0,00
0,02
0,00
0,08
0,42
0,30
0,16
0,02
129,5
133,5
137,5
141,5
145,5
149,5
153,5
mEq
Frequency poligon
Mid point interval
Interval
Determinaciones de sodio

Vall d'Hebron
Institut de Recerca

UNITAT
D'ESTADÍSTICA I
BIOINFORMÀTICA

Steam and Leaf graph

128	106	125	108	98	58	118	92	108	132	32	140	138	96	161
133	104	122	124	110	120	86	115	118	95	83	112	128	127	124
133	115	127	135	89	123	134	94	67	124	155	105	100	112	141
121	112	135	115	64	104	132	98	146	132	93	85	94	116	113
104	115	138	105	144	121	68	107	122	126	88	89	108	115	85
87	88	103	108	109	111	121	124	104	125	102	122	137	110	101
					91	122	138	99	115	104	98	89	119	109

3	2
4	
5	8
6	7 8 4
7	
8	6 3 5 8 9 5 9 7 8 9
9	8 2 6 5 4 8 3 4 1 9 8
10	6 8 5 0 4 7 8 4 2 1 4 9 4 5 3 8 9 8 4
11	8 5 8 2 2 6 3 5 1 0 5 9 5 0 5 2 5
12	8 5 0 7 4 3 4 1 2 6 1 4 2 2 8 5 2 4 7 1
13	2 8 4 2 2 7 8 8 3 3 5 5
14	0 1 6 4
15	5
16	1

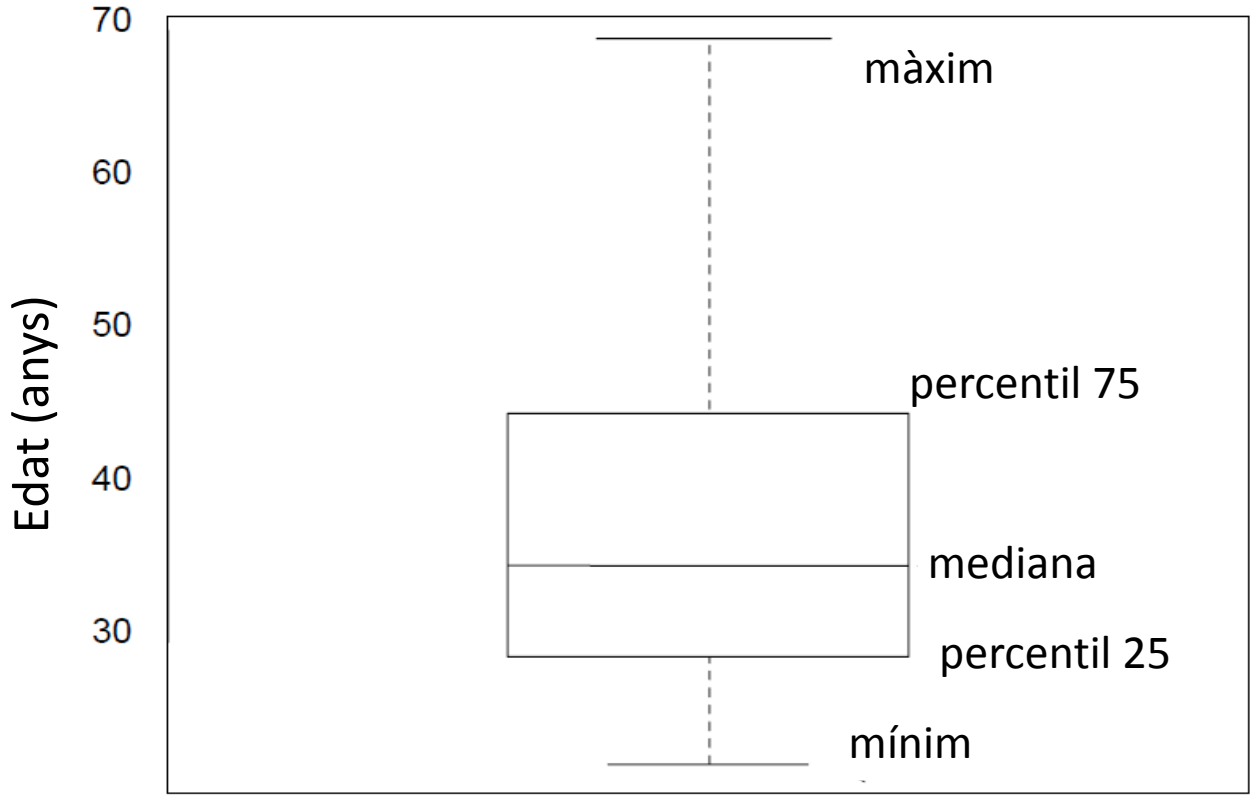
Steam and Leaf graph

128	106	125	108	98	58	118	92	108	132	32	140	138	96	161
133	104	122	124	110	120	86	115	118	95	83	112	128	127	124
133	115	127	135	89	123	134	94	67	124	155	105	100	112	141
121	112	135	115	64	104	132	98	146	132	93	85	94	116	113
104	115	138	105	144	121	68	107	122	126	88	89	108	115	85
87	88	103	108	109	111	121	124	104	125	102	122	137	110	101
					91	122	138	99	115	104	98	89	119	109

3	2
4	
5	8
6	4 7 8
7	
8	3 5 5 6 7 8 8 9 9 9
9	1 2 3 4 4 5 6 8 8 8 9
10	0 1 2 3 4 4 4 4 4 5 5 6 7 8 8 8 8 9 9
11	0 0 1 2 2 2 3 5 5 5 5 5 5 6 8 8 9
12	0 1 1 1 2 2 2 2 3 4 4 4 4 5 5 6 7 7 8 8
13	2 2 2 3 3 5 5 4 7 8 8 8
14	0 1 4 6
15	5
16	1

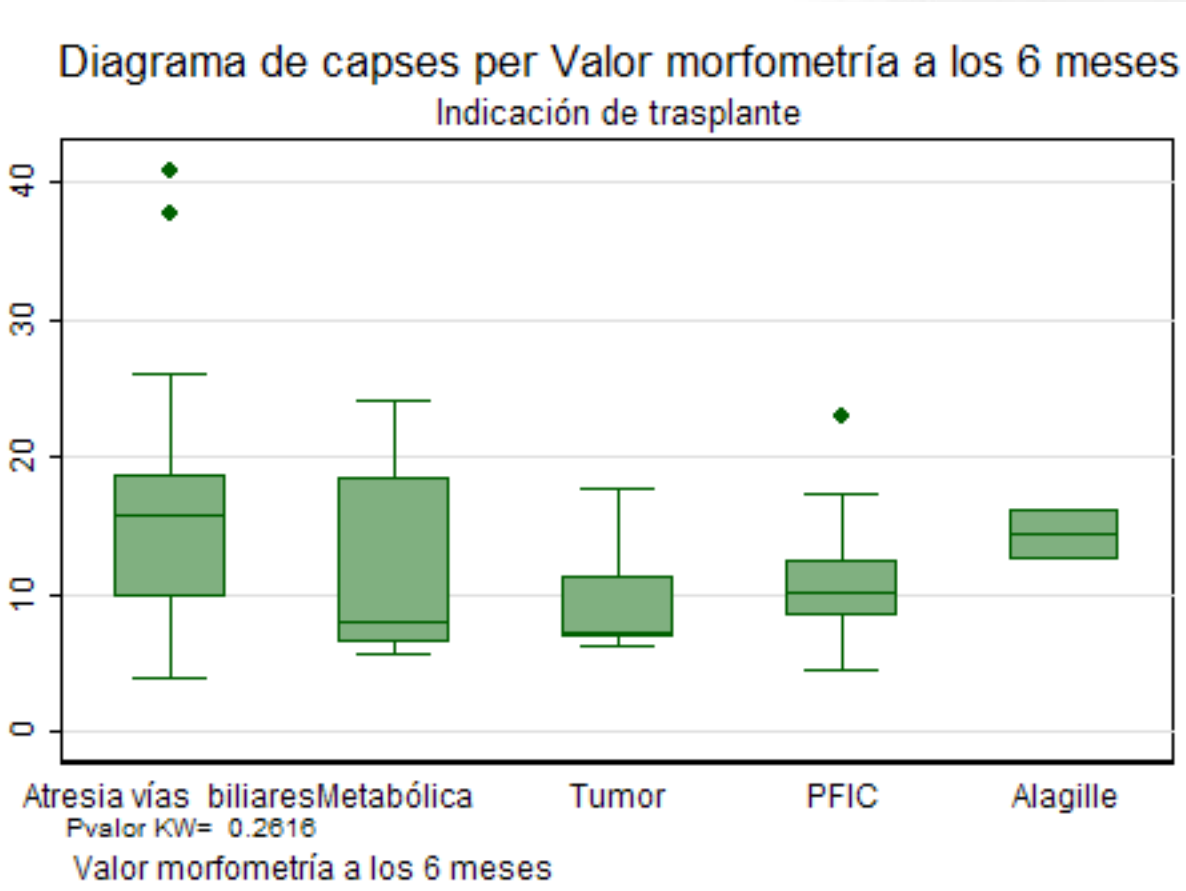
Boxplot

It is graphically represented the “five numbers": box are 25th and 75th percentiles, the middle line is the median (50th percentile) and the ends are the minimum and maximum values.



Boxplot

The boxplot is a quick way to identify outliers in the sample (they can not be "outliers")



Index