

Descriptive Statistics. Summary and Graphs

Curs d'Estadística Bàsica per a la Recerca Biomèdica

UEB – VHIR

Miriam Mota-Foix Santiago Pérez-Hoyos

miriam.mota@vhir.org santi.perezhoyos@vhir.org

-
- 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

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

Index

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

- **Population:** The population 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 elected).
- **Variable:** Feature measurable or observable that represents the concepts 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 ANALYSIS

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 in 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

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 & exercicis

TYPE 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: N° of admissions, N° of programmed visits

CONTINUOUS

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**
 - Are those that answer the research question
- **Explain, independent or exposure variables**
 - 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, to transform variables and detect outliers

Index

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

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 interval(IQR)
- Shape
 - Asymmetry
 - Kurtosis

Location measures



Mean

Median

Mode

Mean

μ

- 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 divided in **two parts** the observations
- Observations are ordered from lowest to higheest 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 de Classe

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
```

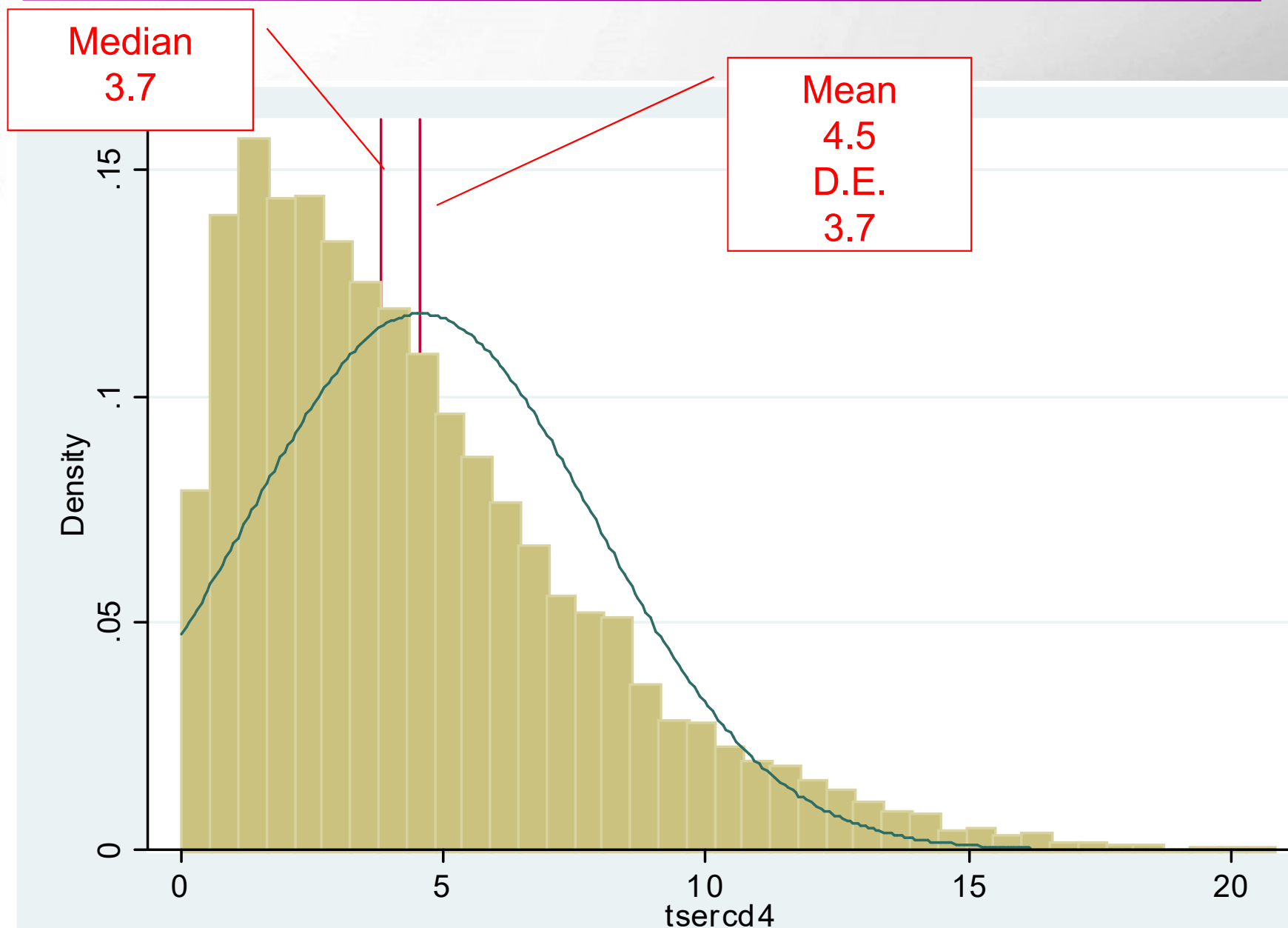
```
> median(x1)
```

```
[1] 6
```

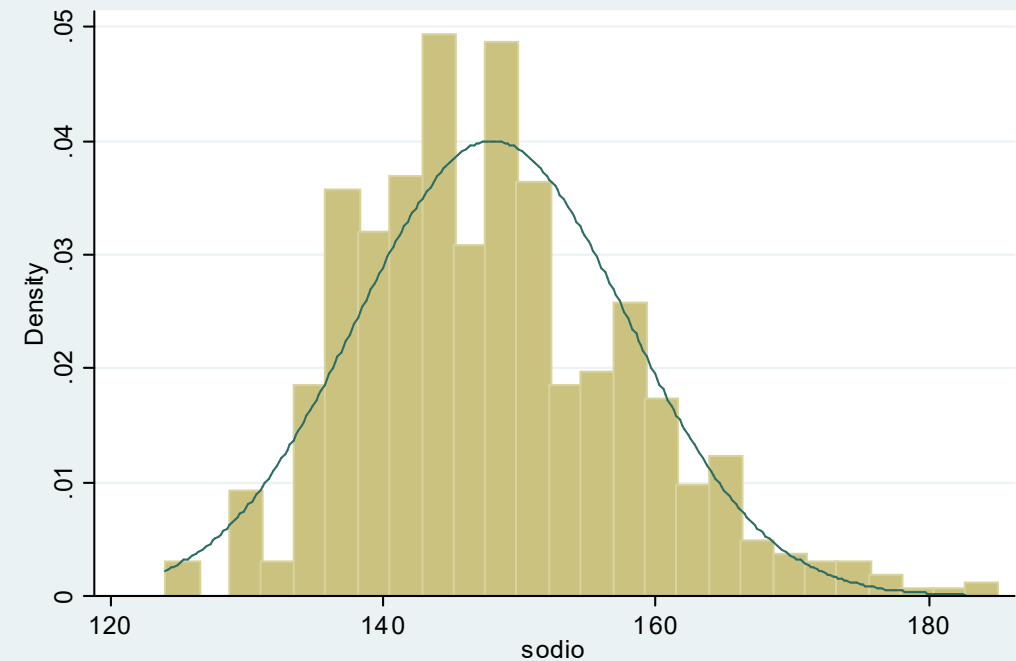
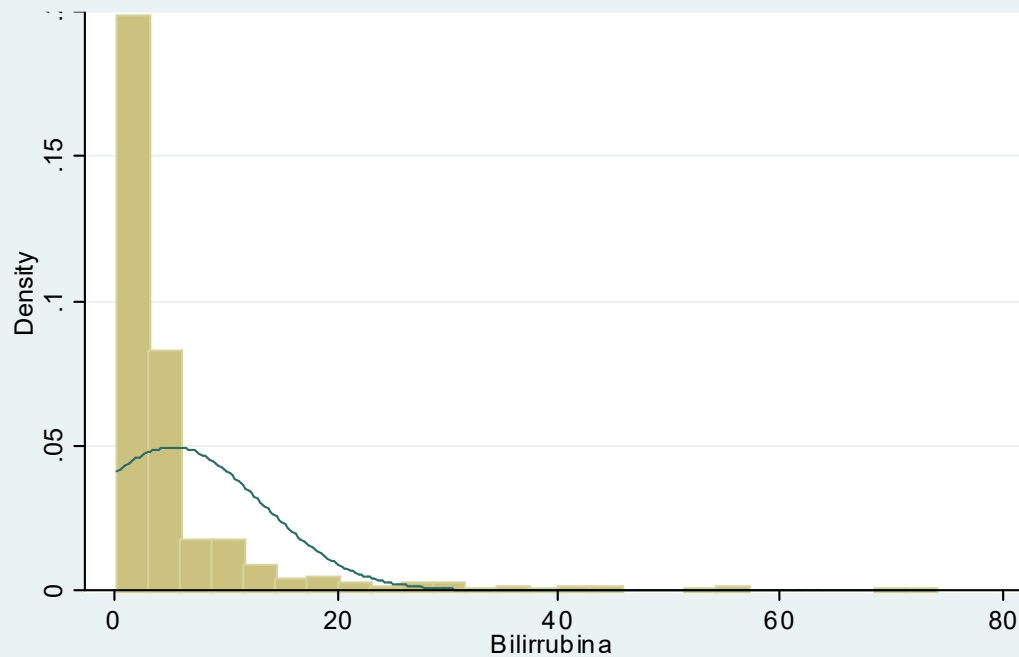
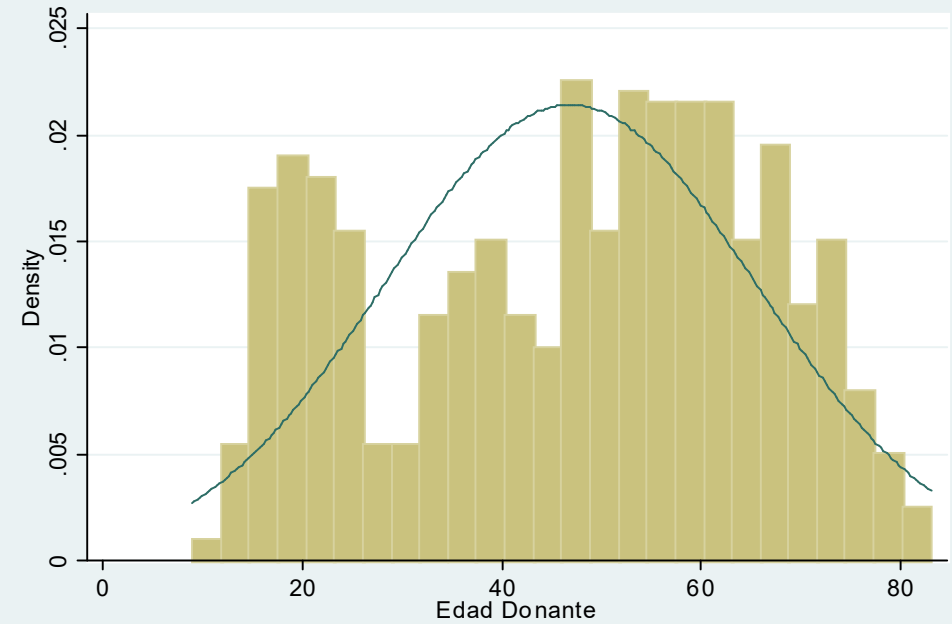
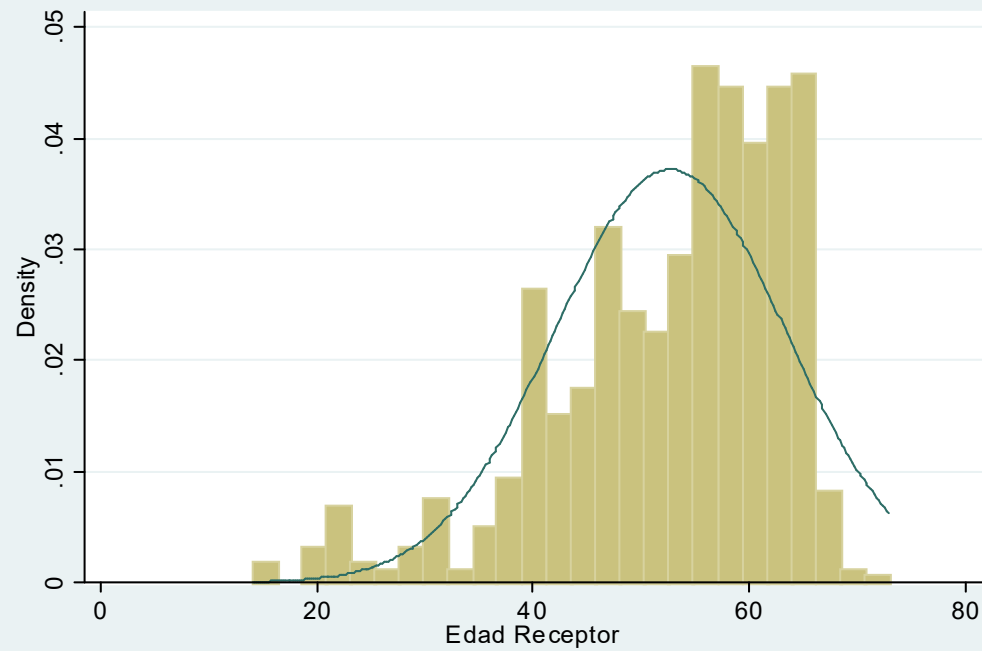
```
> median(x2)
```

```
[1] 6
```

Mean or Median



Transplant study (real data)



- 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



Dispersion or variability measures



Range (Maximum-Minimum)

Variance

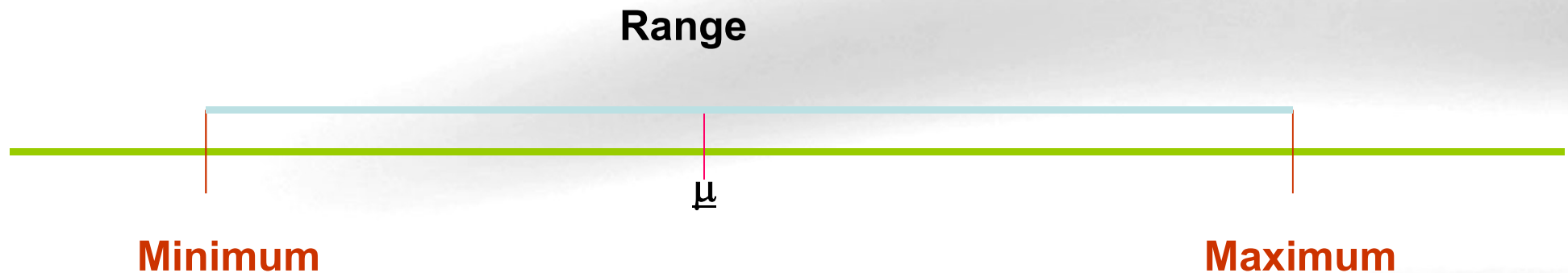
Standard Deviation

Variation Coefficient

Percentile

Interquartilic Interval (IQR)

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



Example Stay days

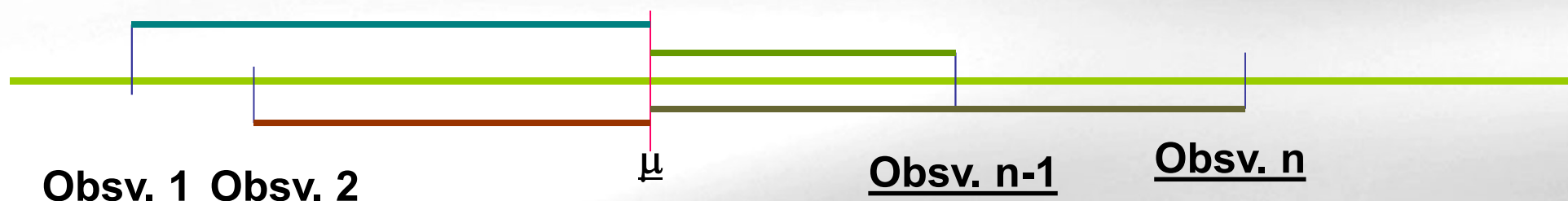
3, 4, 6, 9, 12

$$\text{Range} = 12 - 3 = 9$$

3, 4, 6, 9, 20

$$\text{Range} = 20 - 3 = 17$$

- 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
- Is measured in the same units of variable

Example Stay days

3, 4, 6, 9, 12

Variance =10.96

Std. Dev.=3.31

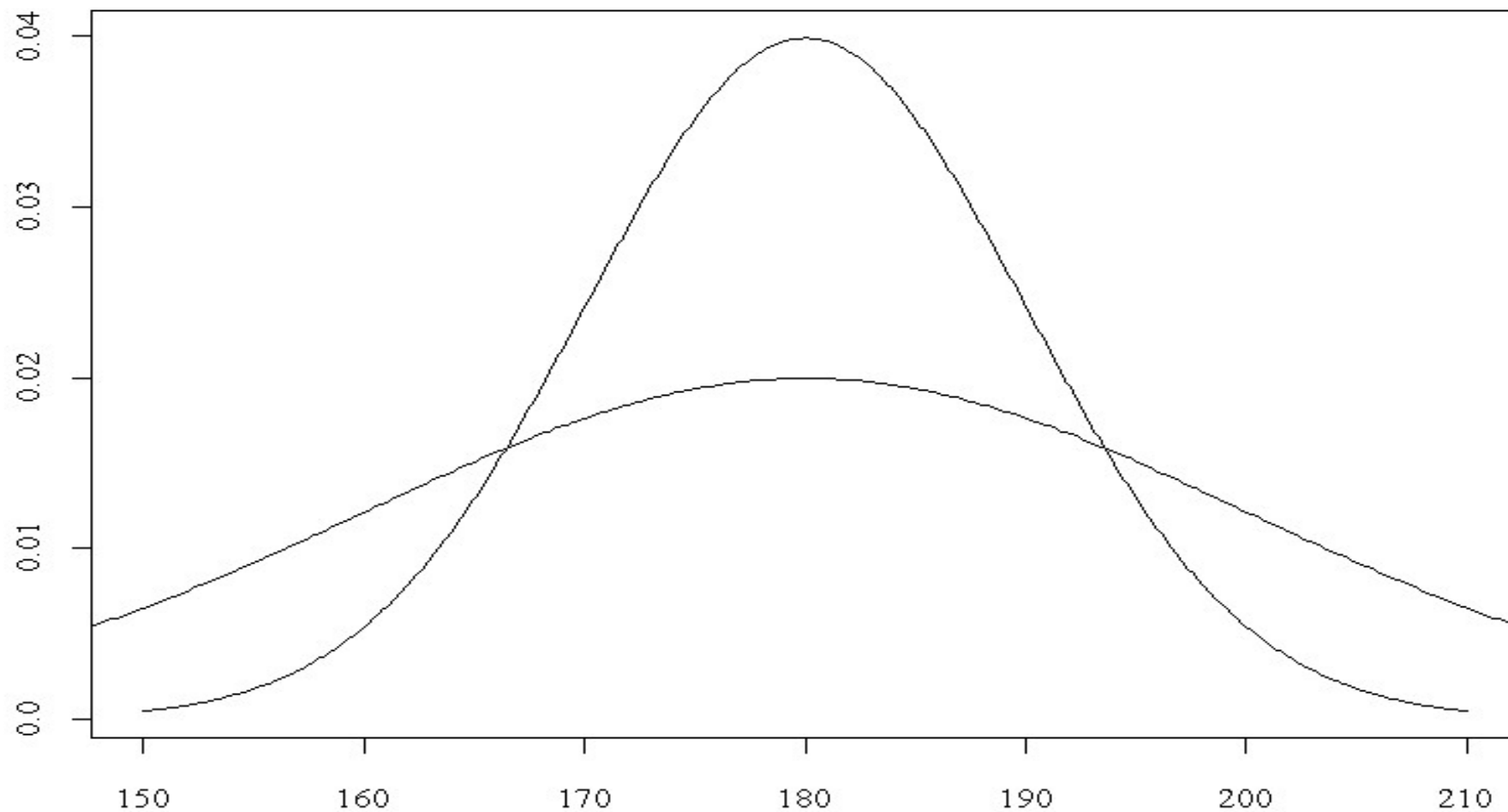
3, 4, 6, 9, 20

Variance= 37.84

Std. Dev.= 6.15

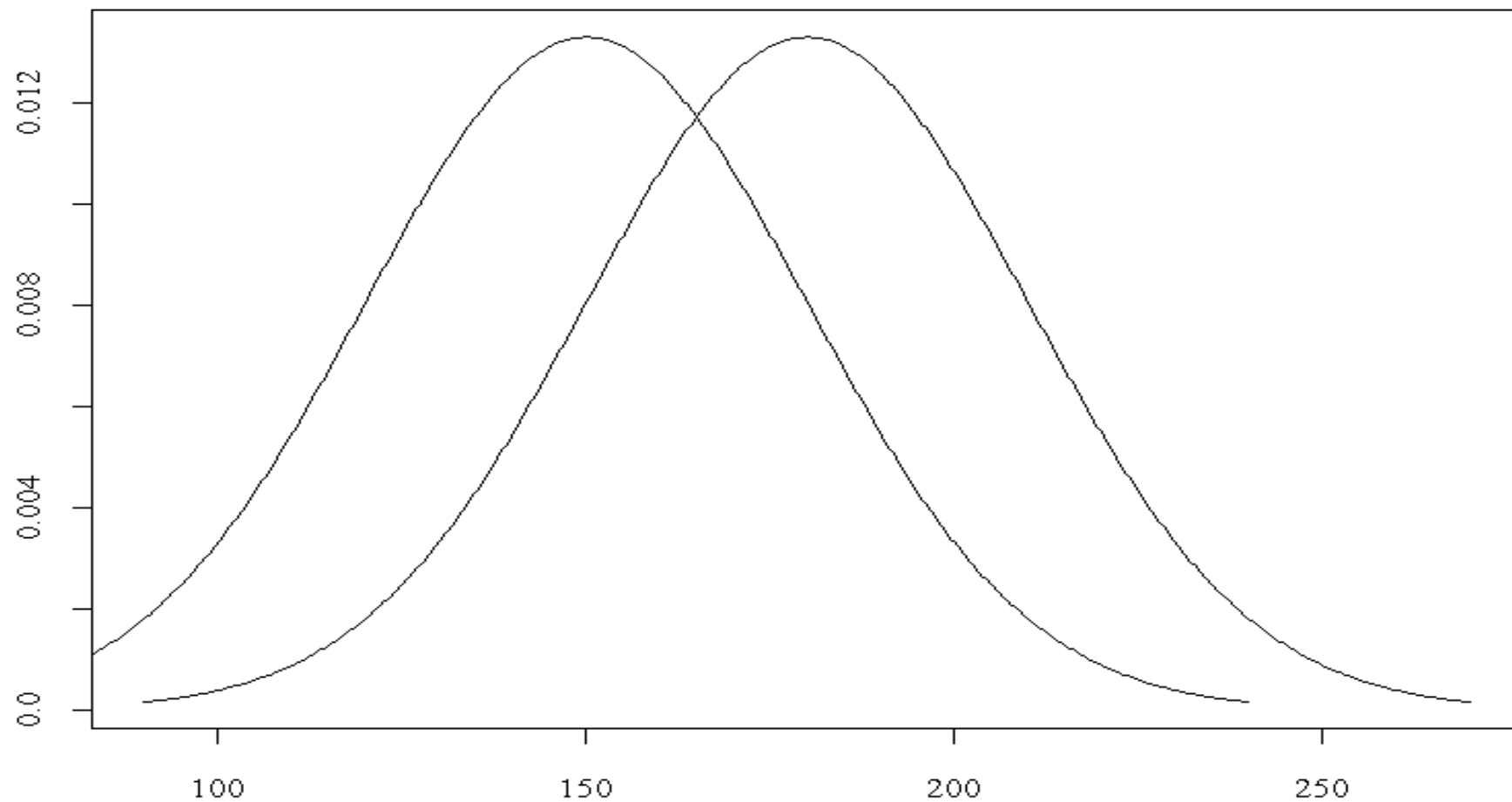
Same mean , different variances

Mismas medias, diferentes varianzas



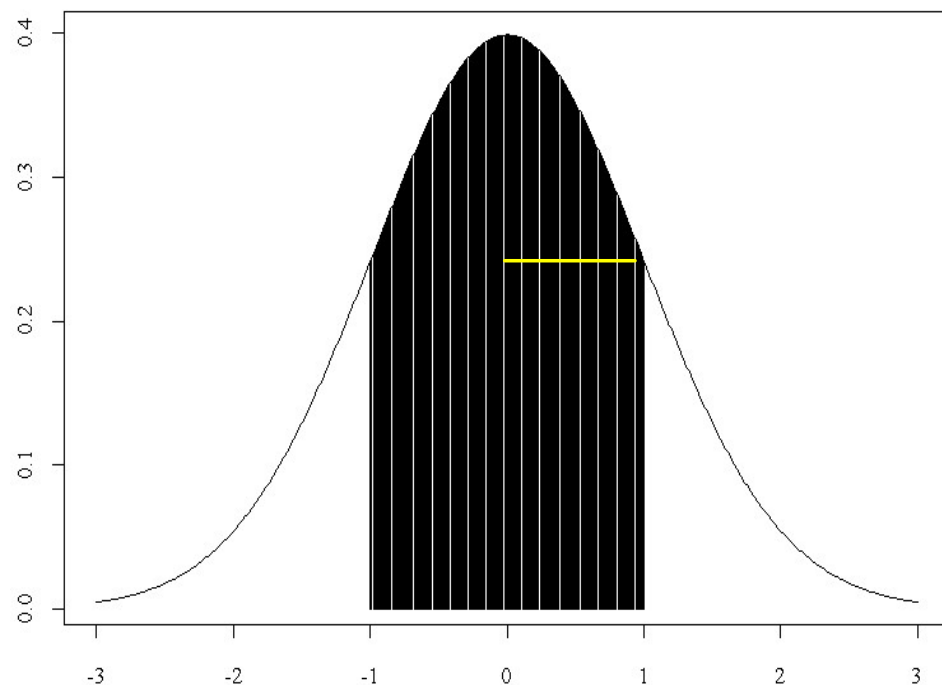
Same variances, different means

Mismas varianzas, diferentes medias



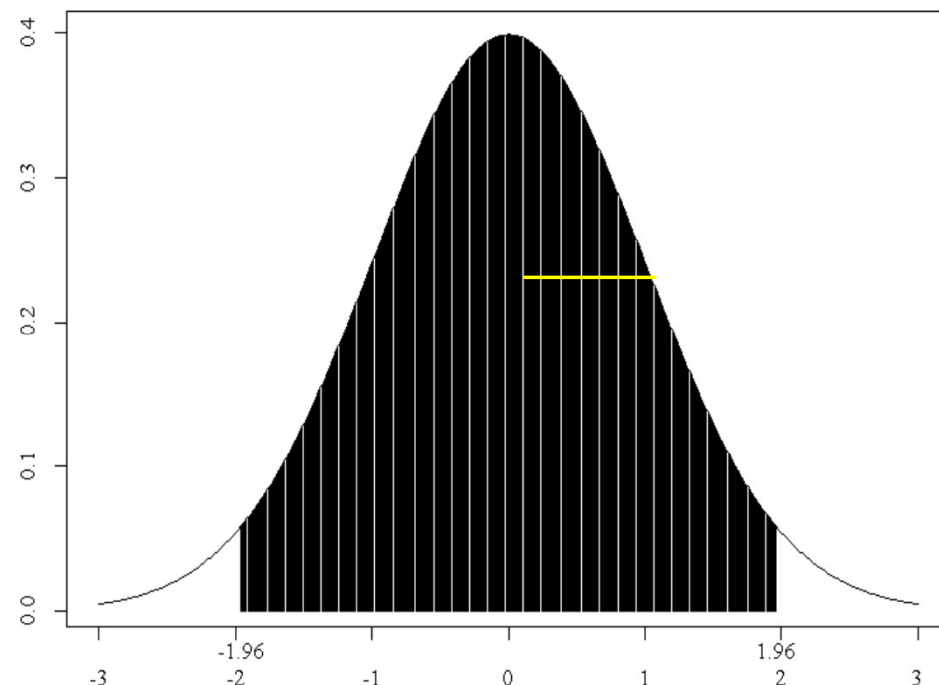
Normal distribution

Densidad Z Normal(0,1)



Probabilidad 68%

Densidad Z Normal(0,1)

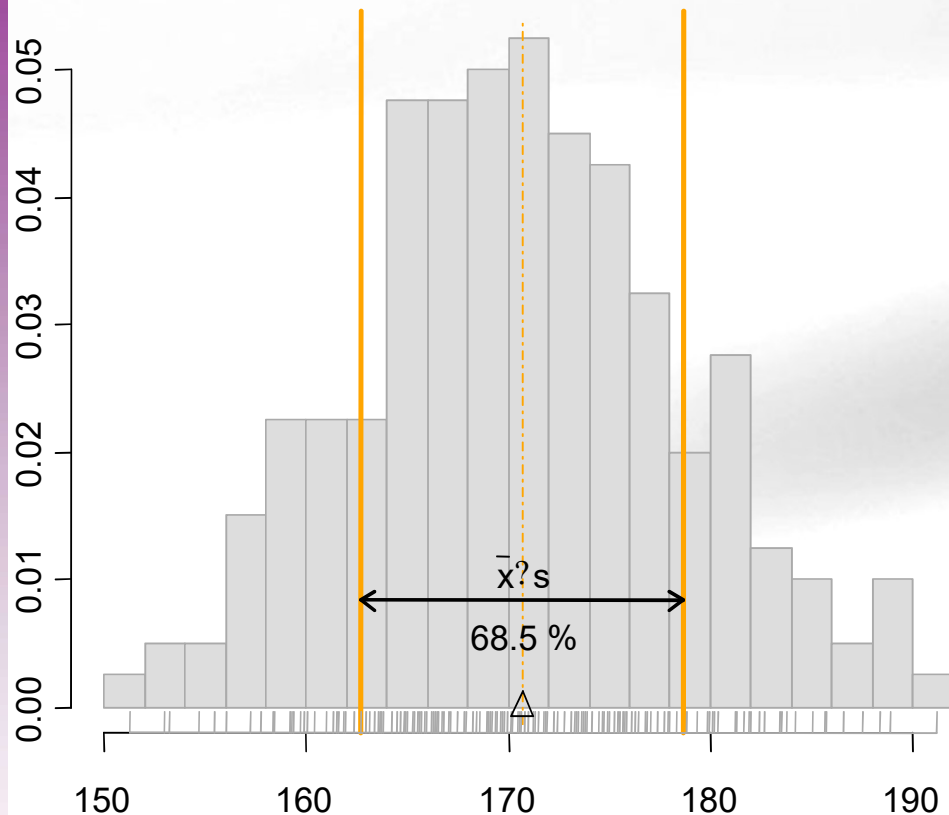


Probabilidad 95%

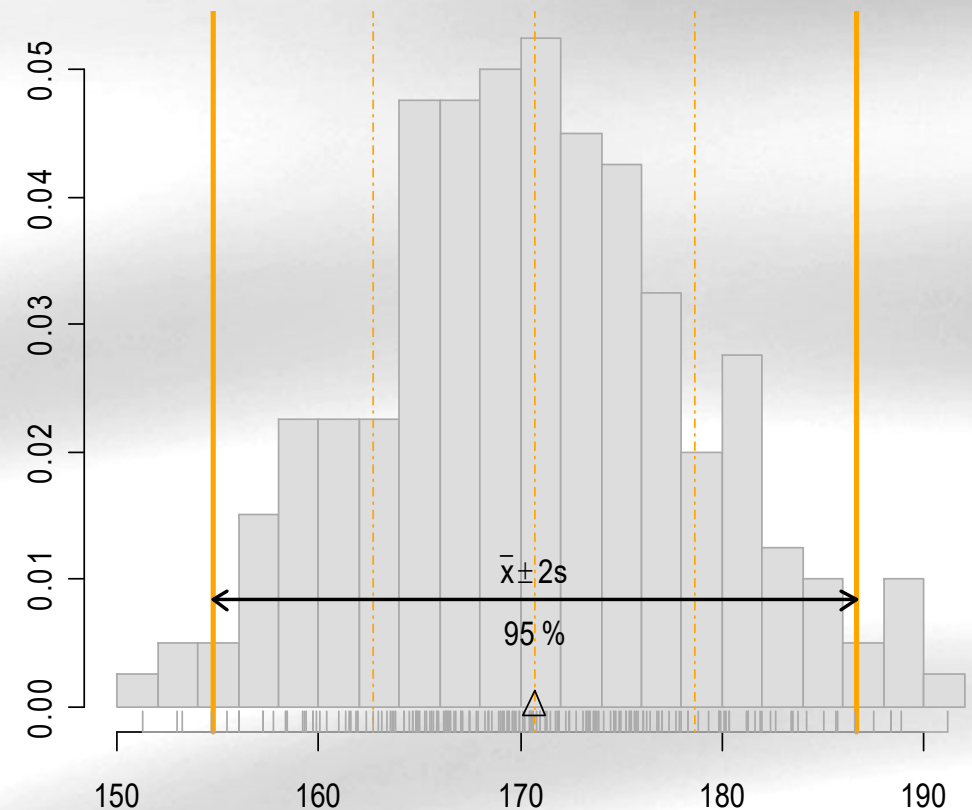
- ✓ Centered in the mean . Between 1 SD ther are about 68% of the observations.

- ✓ Between two SD is about 95% of observations

Symmetric distribution of data

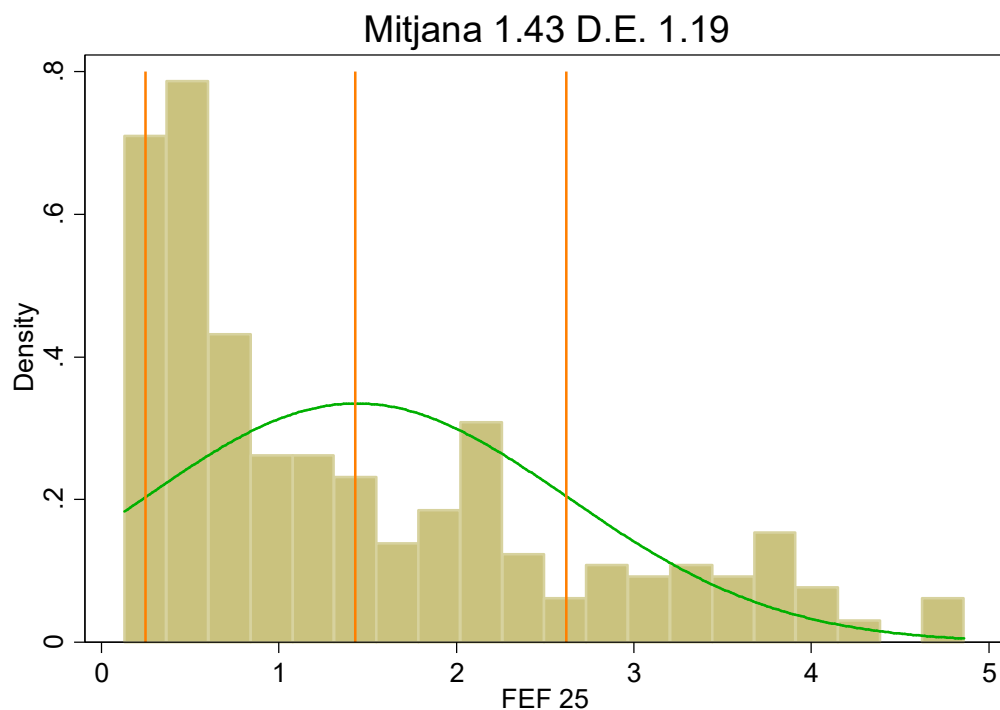


- ✓ Centered in the mean . Between 1 SD ther are about 68% of the observations.

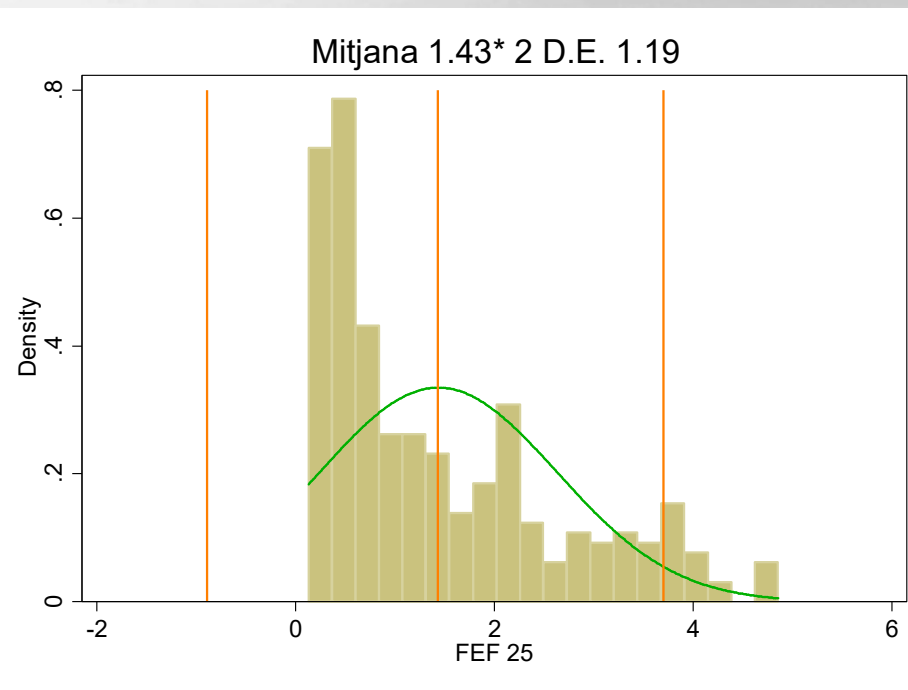


- ✓ Between two SD is about 95% of observations

Assymetric distribution



- ✓ Centered in the mean . Between 1 SD the % of observations is not the 68%



- ✓ Between two SD the % of observations is not the 95%

Variation coefficient

- 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

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

$$\left\{ \begin{array}{l} X = \text{Kg} \\ s = \end{array} \right.$$

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

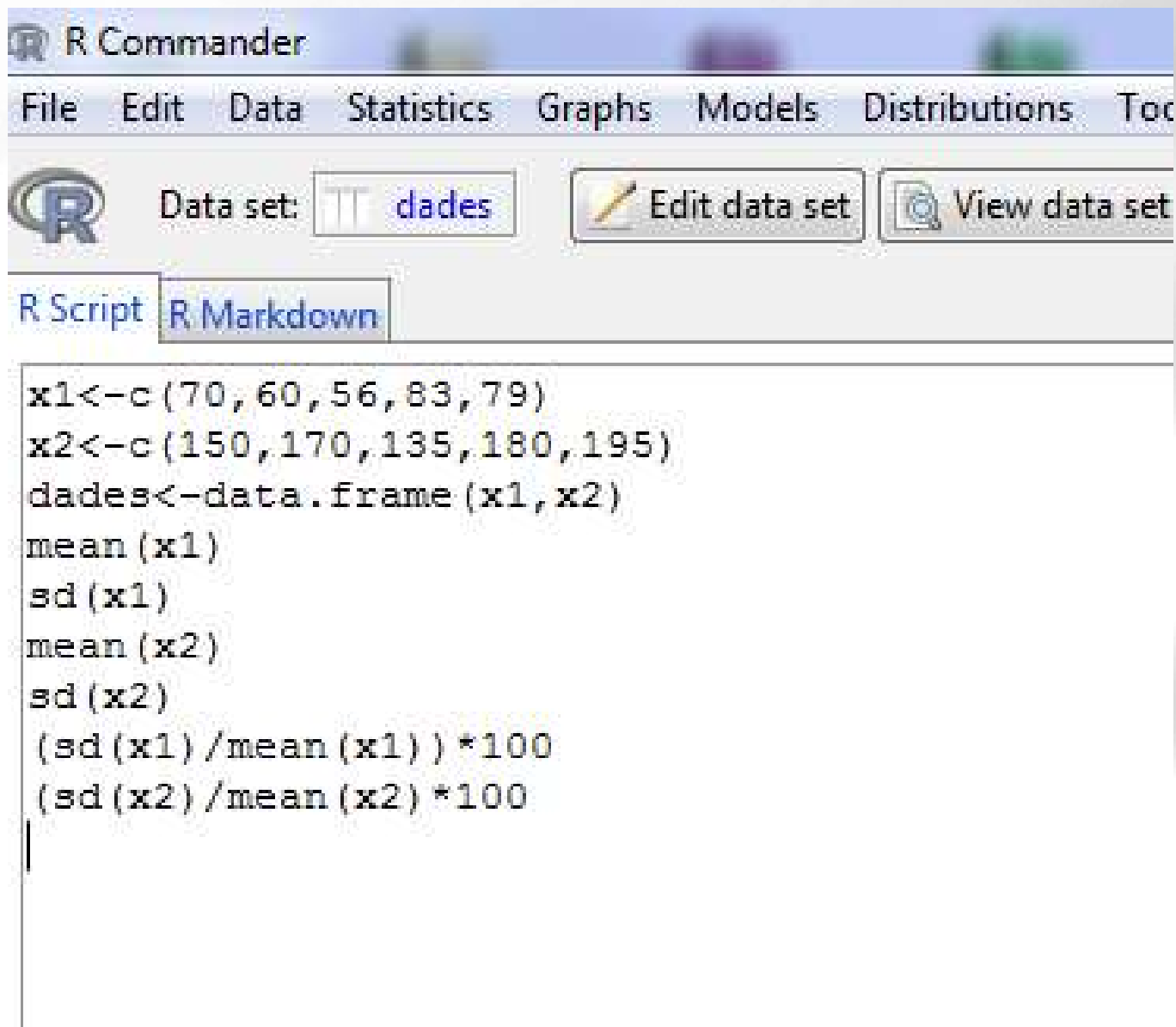
$$\left\{ \begin{array}{l} X = \\ s = \end{array} \right.$$

Which variable have more variation? Weight or Pression



Calcualte CV

$$\left\{ \begin{array}{l} CV_{WGT} = \\ CV_{BP} = \% \end{array} \right.$$



The image shows a screenshot of the R Commander application window. The title bar reads "R Commander". The menu bar includes "File", "Edit", "Data", "Statistics", "Graphs", "Models", "Distributions", and "Tools". Below the menu bar, there is a toolbar with the R logo, a "Data set:" label, a text box containing "dades", and two buttons: "Edit data set" and "View data set". Below the toolbar, there are two tabs: "R Script" (which is active) and "R Markdown". The main text area contains the following R code:

```
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

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

$$\bar{X} = 69.6 \text{ Kg}$$

$$s = 11.67$$

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

$$\bar{X} = 166 \text{ mmHg}$$

$$s = 23.82$$

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

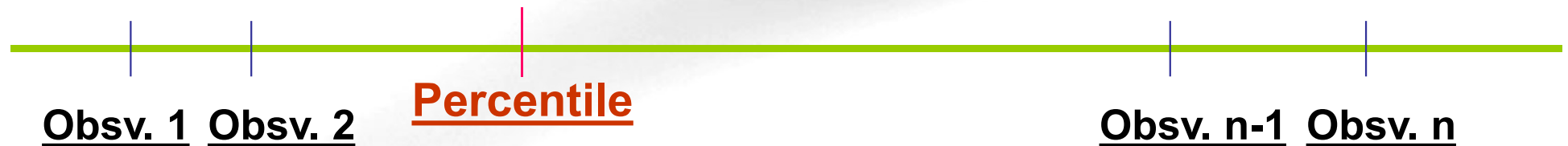


Calculem CV

$$CV_{\text{PES}} = 16.7\%$$

$$CV_{\text{TA}} = 14.3\%$$

- 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 altered by extreme observations
 - Interquartile range is difference between 25 and 75 percentile



Summary measures in R Commander

The image shows the R Commander interface. The 'Statistics' menu is open, displaying a list of options: Summaries, Contingency tables, Means, Proportions, Variances, Nonparametric tests, Dimensional analysis, and Fit models. The 'Summaries' option is selected, which has opened a sub-menu. In this sub-menu, 'Numerical summaries...' is highlighted. Other options in the sub-menu include 'Active data set', 'Frequency distributions...', 'Count missing observations', 'Table of statistics...', 'Correlation matrix...', 'Correlation test...', and 'Shapiro-Wilk test of normality...'. In the background, the 'Numerical Summaries' dialog box is visible. It has two tabs: 'Data' and 'Statistics'. The 'Statistics' tab is active, showing a list of variables: 'bmi', 'dbp', 'edat', 'edatdiag', 'numpacie', and 'sbp'. The 'edat' variable is selected. Below the variable list is a button labeled 'Summarize by groups...'. At the bottom of the dialog, there are three buttons: 'OK' (with a green checkmark), 'Cancel' (with a red X), and 'Apply' (with a green circular arrow). In the foreground, another 'Numerical Summaries' dialog box is shown, but it is partially obscured. This dialog has the 'Data' tab selected. It contains a list of statistical measures with checkboxes: 'Mean' (checked), 'Standard Deviation' (checked), 'Standard Error of Mean' (unchecked), 'Interquartile Range' (checked), 'Coefficient of Variation' (checked), 'Skewness' (unchecked, with radio buttons for Type 1, Type 2, and Type 3), 'Kurtosis' (unchecked, with radio buttons for Type 2 and Type 3), and 'Quantiles' (checked, with a text box containing '0, .25, .5, .75, 1'). At the bottom of this dialog are buttons for 'Help', 'Reset', 'OK', 'Cancel', and 'Apply'.

Statistics | Graphs | Models | Distributions | Tools | Help

- Summaries ▸
 - Active data set
 - Numerical summaries...
 - Frequency distributions...
 - Count missing observations
 - Table of statistics...
 - Correlation matrix...
 - Correlation test...
 - Shapiro-Wilk test of normality...
- Contingency tables ▸
- Means ▸
- Proportions ▸
- Variances ▸
- Nonparametric tests ▸
- Dimensional analysis ▸
- Fit models ▸

Numerical Summaries

Data | Statistics

Variables (pick one or more)

- bmi
- dbp
- edat
- edatdiag
- numpacie
- sbp

Summarize by groups...

Numerical Summaries

Data | Statistics

- ☒ Mean
- ☒ Standard Deviation
- ☐ Standard Error of Mean
- ☒ Interquartile Range
- ☒ Coefficient of Variation
- ☐ Skewness ☐ Type 1
- ☐ Kurtosis ☒ Type 2 ☐ Type 3
- ☒ Quantiles: 0, .25, .5, .75, 1

Help Reset OK Cancel Apply

Summary measures in R Commander

The screenshot shows the R Commander interface with the 'Numerical Summaries' window open. The window displays a table of statistical measures and their values for a dataset. The measures are: mean, sd, IQR, cv, 0%, 25%, 50%, 75%, 100%, and n. The values are: 52.16779, 11.77285, 17, 0.2256728, 31, 43, 50, 60, 86, and 149. Three callouts point to specific values: 'Minimum' points to 31 (under cv), 'Median' points to 50 (under 50%), and 'Maximum' points to 86 (under 100%). Below the table, there are buttons for 'Cancel' and 'Apply'. At the bottom, there is a detailed view of the 'Statistics' tab, showing checkboxes for Mean, Standard Deviation, Standard Error of Mean, Interquartile Range, Coefficient of Variation, Skewness (Type 1), Kurtosis (Type 2), and Quantiles (0, .25, .5, .75, 1). The 'OK' button is highlighted.

mean	sd	IQR	cv	0%	25%	50%	75%	100%	n
52.16779	11.77285	17	0.2256728	31	43	50	60	86	149

Minimum

Median

Maximum

Cancel Apply

Statistics

☒ Mean

☒ Standard Deviation

☐ Standard Error of Mean

☒ Interquartile Range

☒ Coefficient of Variation

☐ Skewness ☐ Type 1

☐ Kurtosis ☒ Type 2

☐ Type 3

☒ Quantiles: 0, .25, .5, .75, 1

Help Reset OK Cancel Apply

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

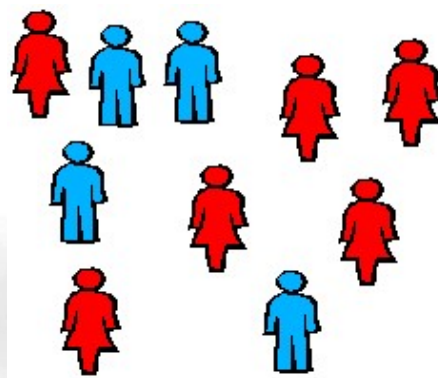
1. Contingency Tables

2. Graphs

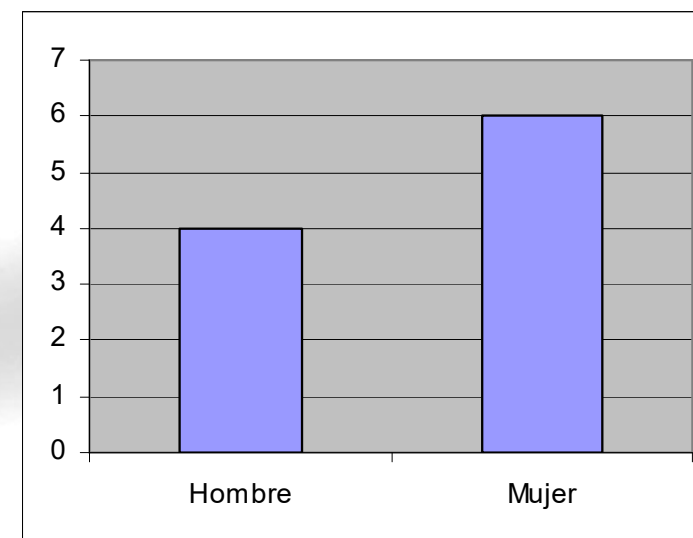
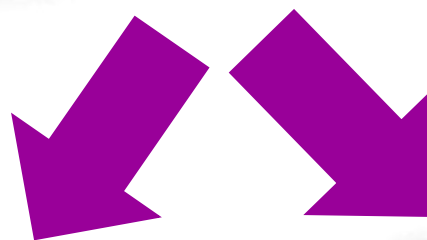
5. Examples & exercises

Summary of variables

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



Género	Frec.
Hombre	4
Mujer	6



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

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 Freqüència

Cate
go
ries

Nº of subjects by
category

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

Percentage of subjects
Freq / Total

Nº of subjects by
category

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

Frequency table

Percentage of subjects
Freq / Total

Cate
go
ries

Nº of subjects by
category

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			500	1'00
TOTAL			500	1'00

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

Frequency table

Percentage of subjects
Freq / Total

Nº of subjects by
category

Cate
go
ries

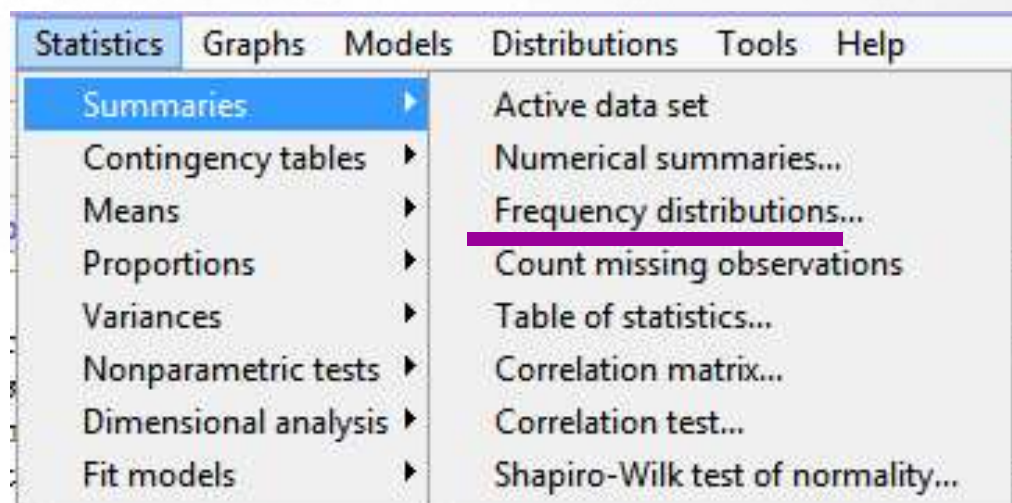
NÚMERO DE HIJOS ENTRE MUJERES DE 20 Y 30 AÑOS

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			500	
TOTAL			500	

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

Accumulated
Frequency up to
category
Freq Abs/Total

Frequency tables in R Commander



counts:

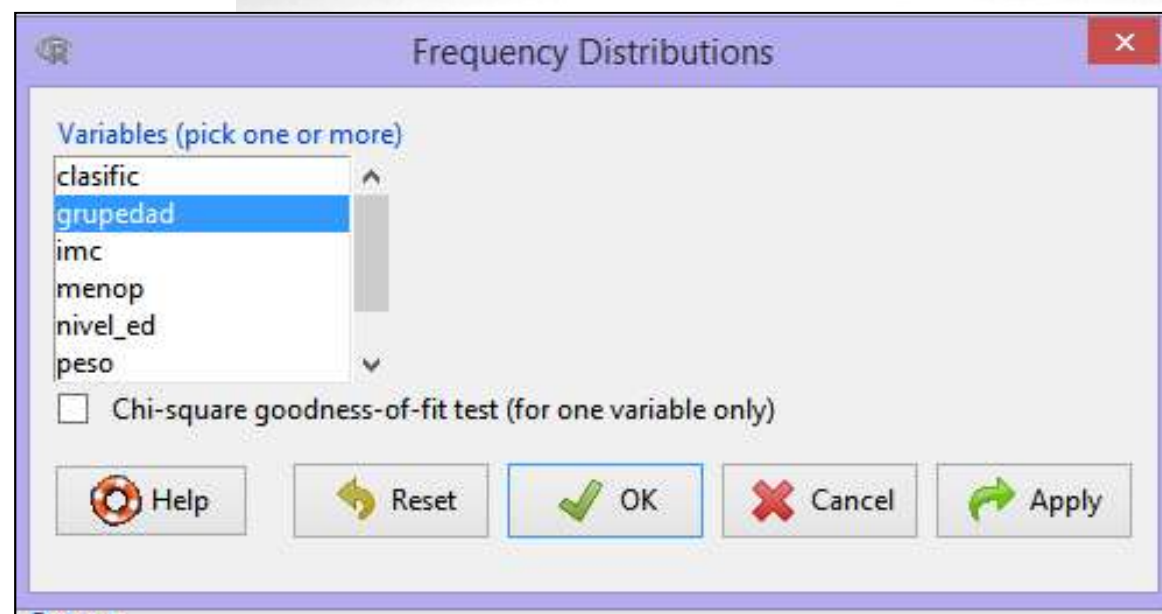
grupedad

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

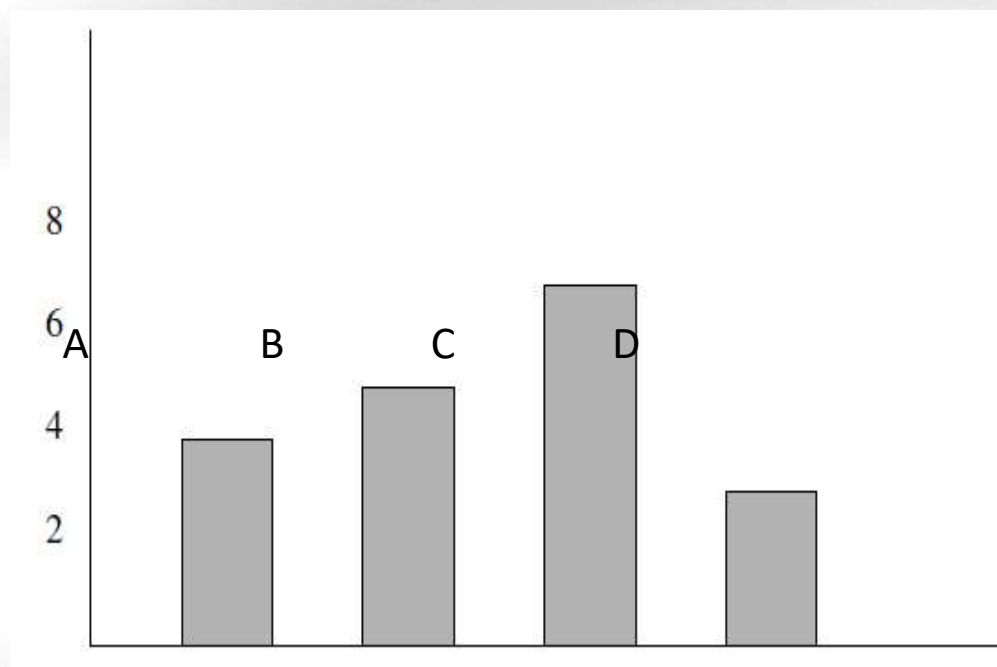
PESO

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

Bar Graph

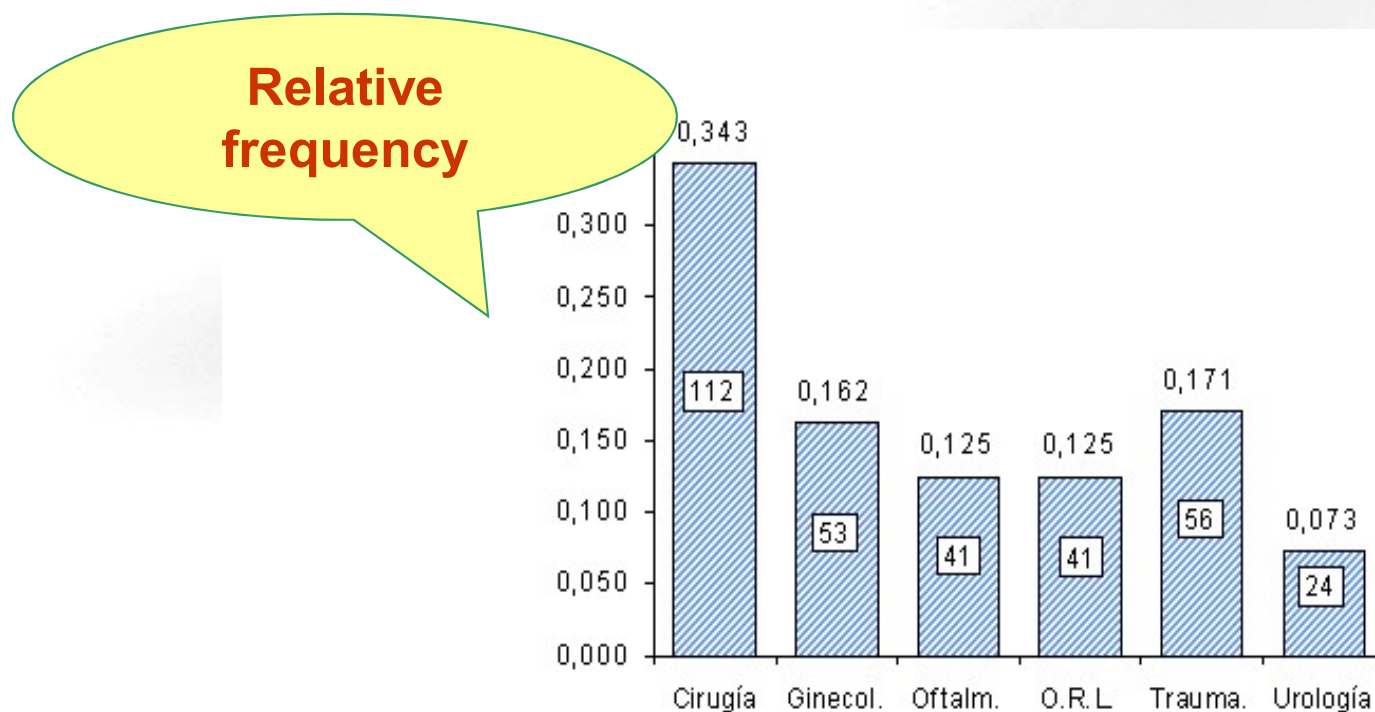
Categories are represented in X axis and frequencies in Y axis

frequències



Bar Graph

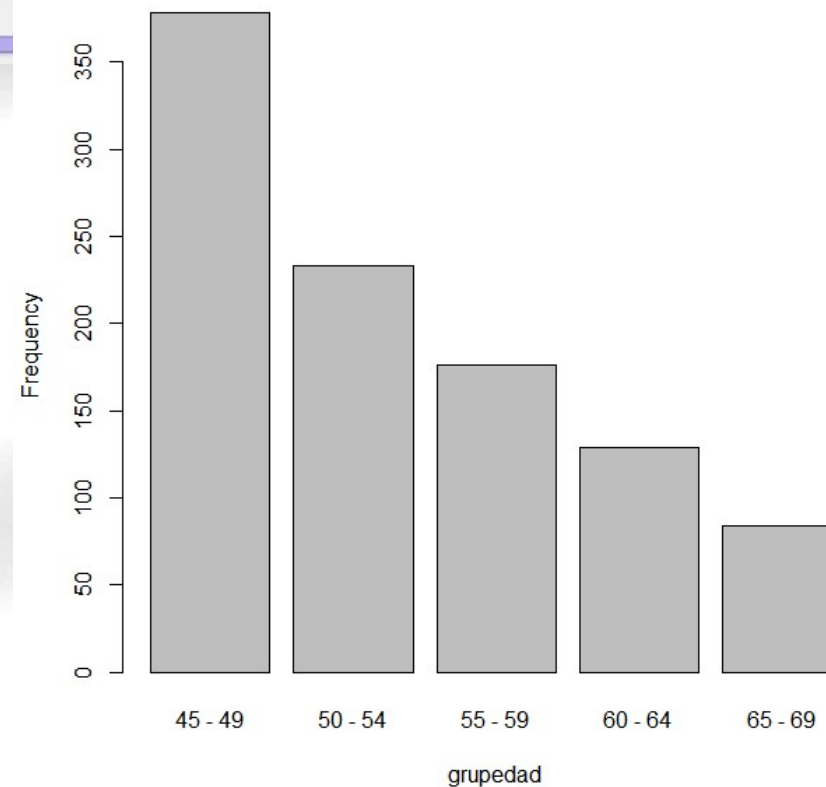
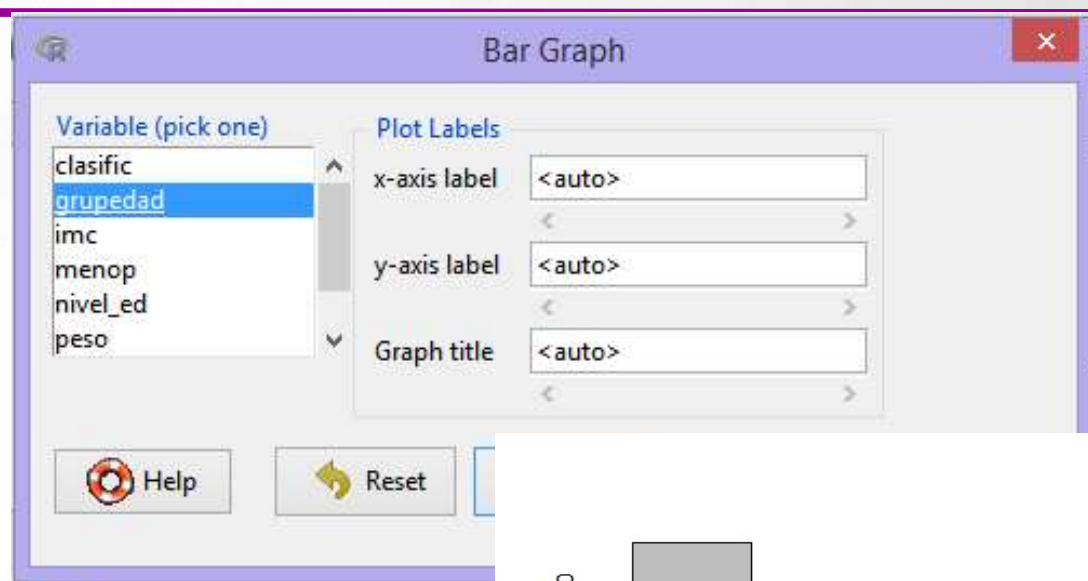
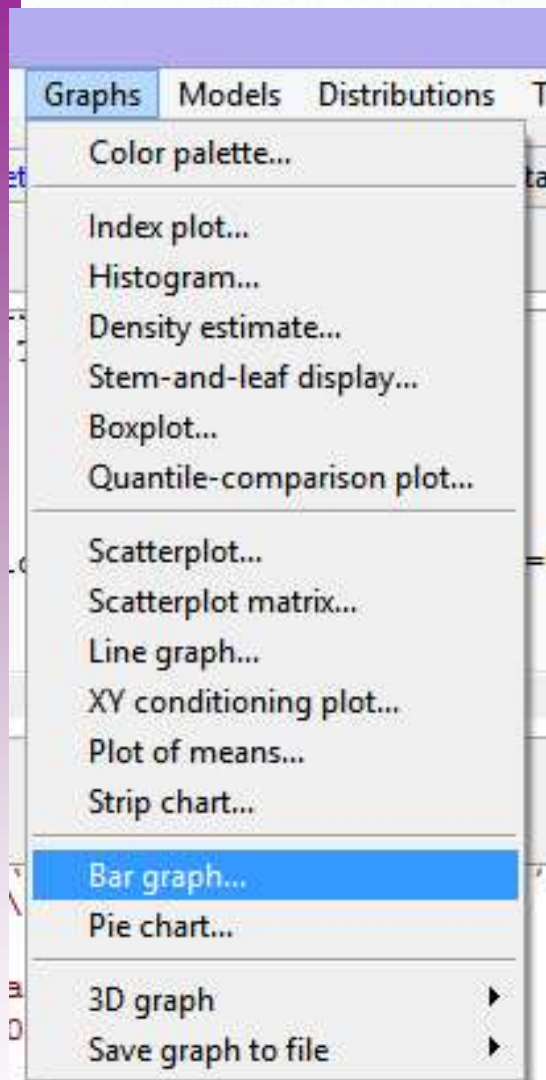
For comparing two population better use relative frequencies



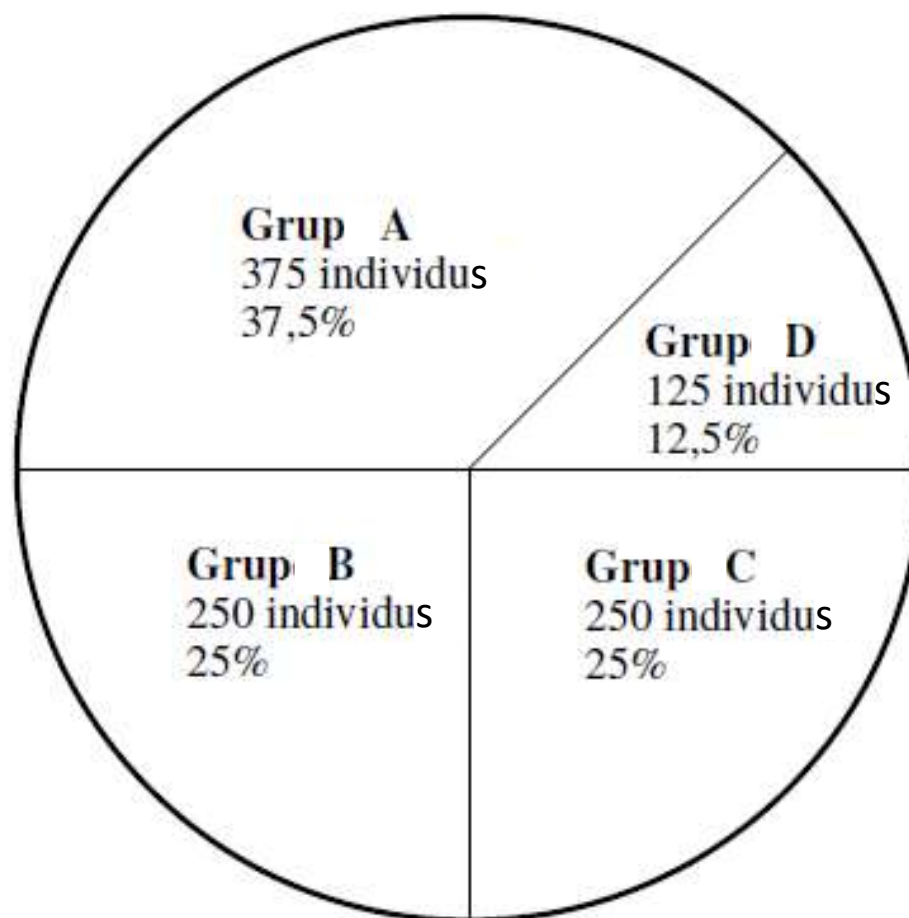
Intervenciones quirúrgicas

Categories

Graphs in R Commander



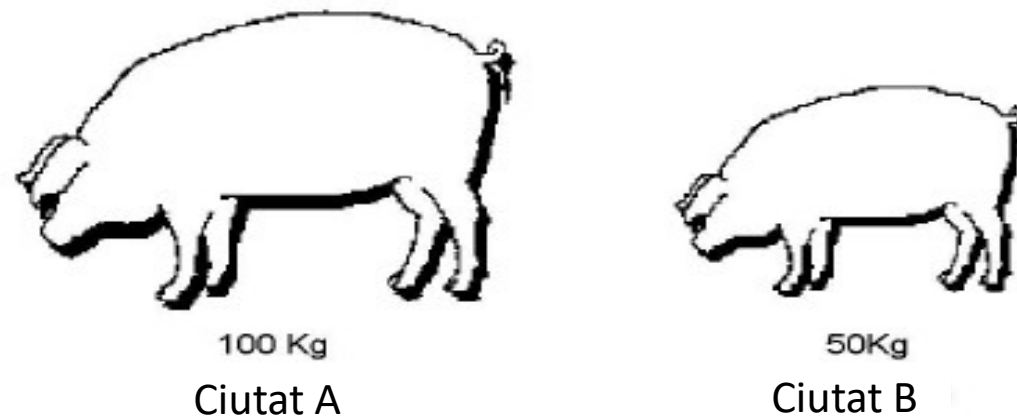
Pie Graph



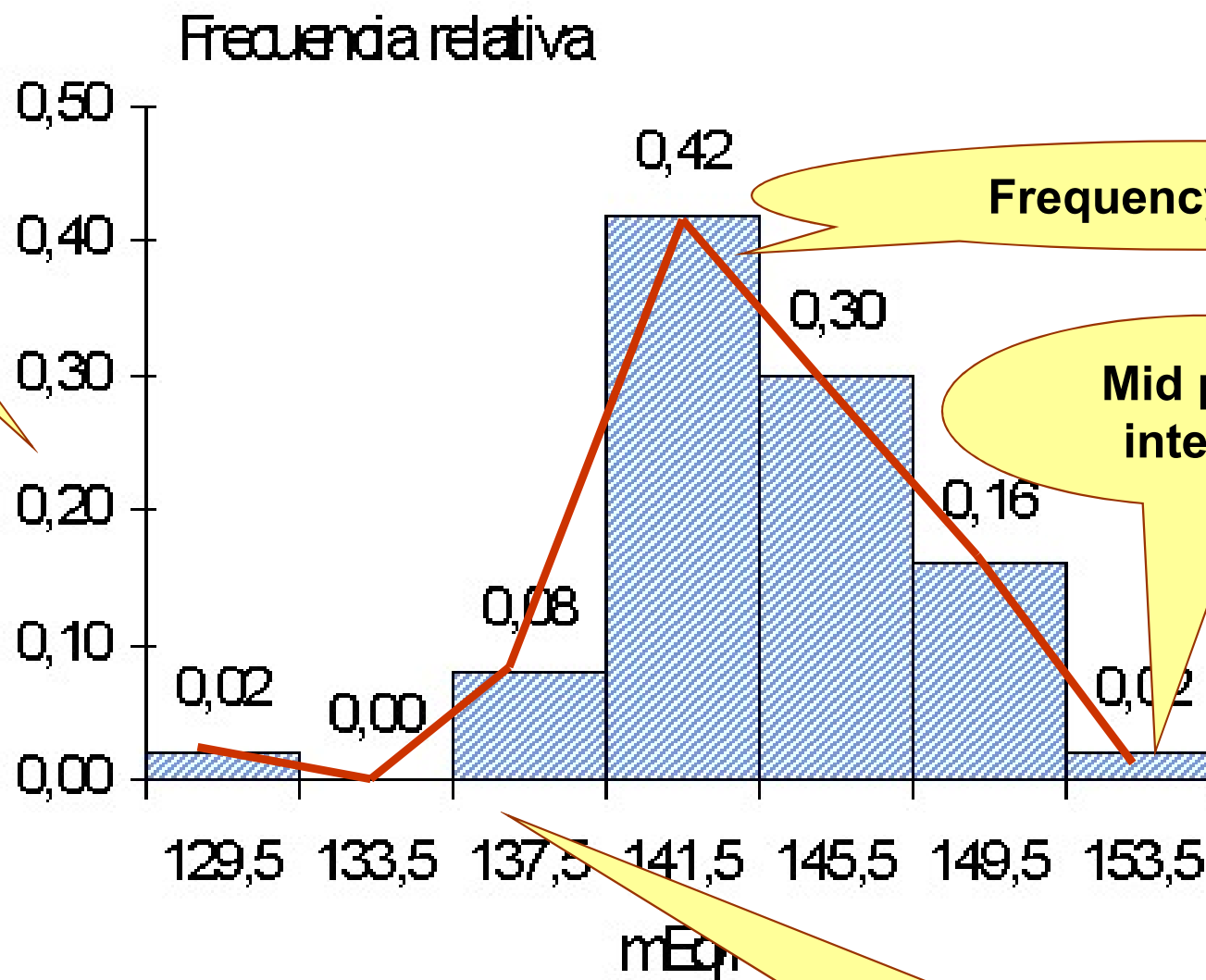
Pictograms

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



Histogram



Determinaciones de sodio

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

```

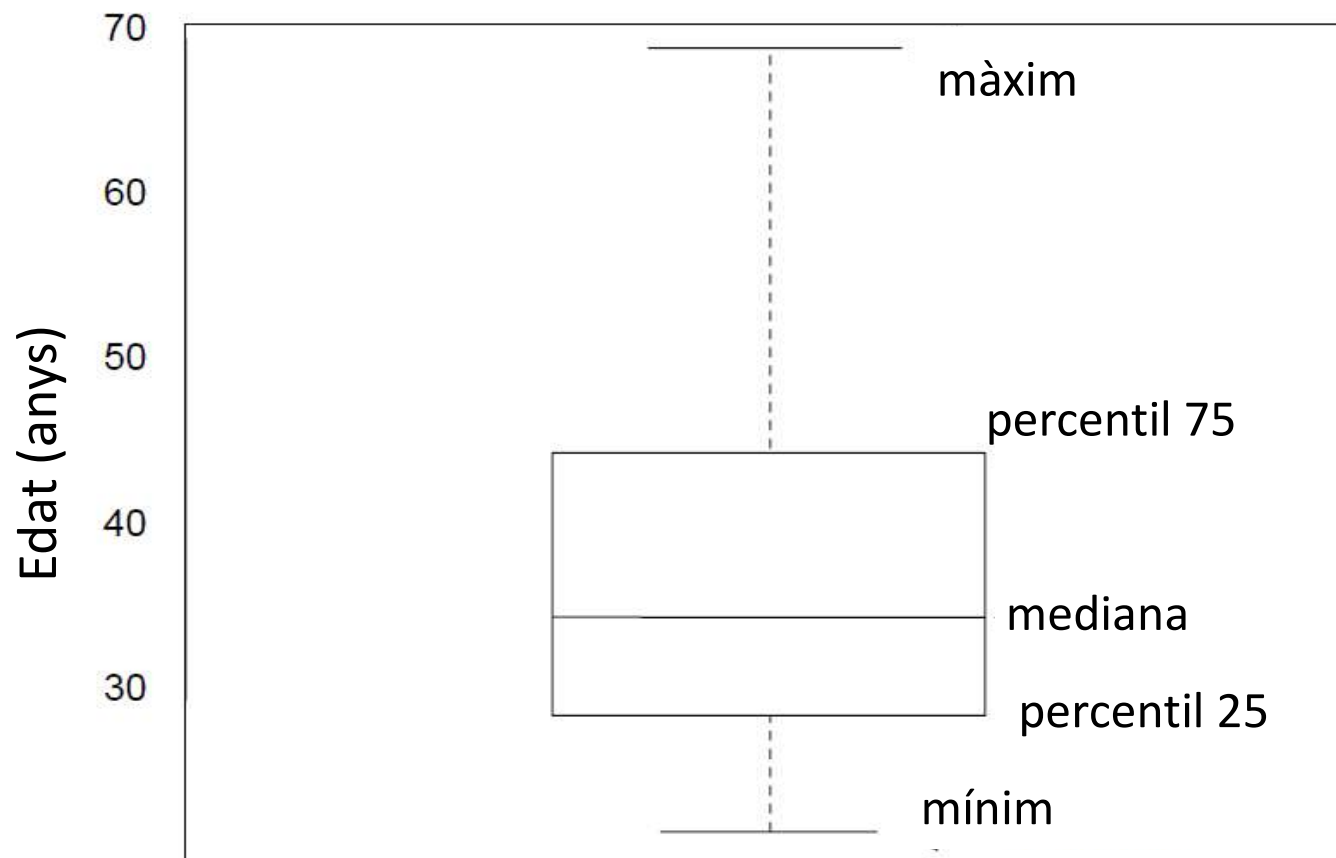

Steam and Leaf graph

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

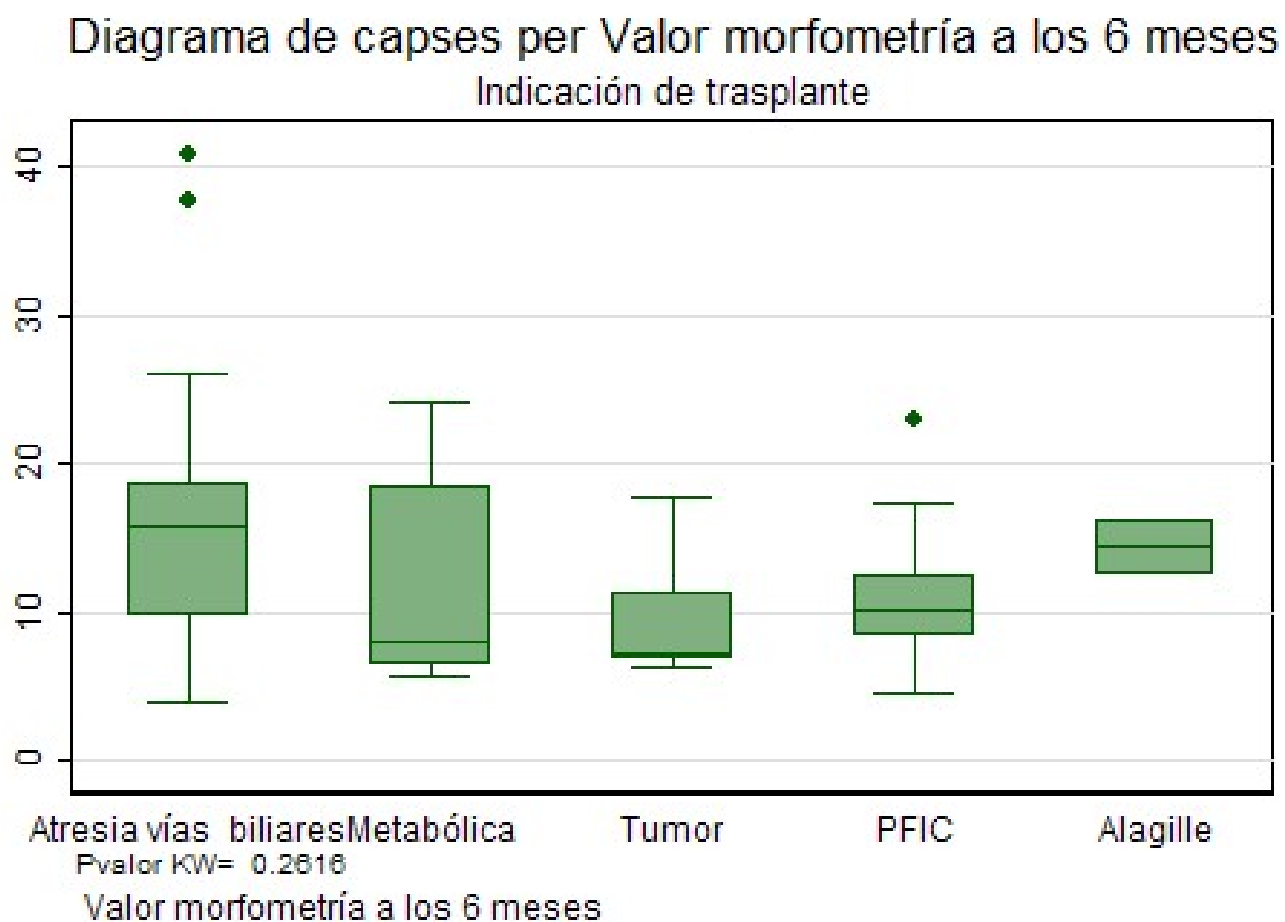
Boxplot

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

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