Applied Biostatistics with R and rk.Teaching

Santiago Angulo Díaz-Parreño (sangulo@ceu.es)
Juan Carlos Garro Garro (garro.eps@ceu.es)
Eduardo López Ramírez (elopez@ceu.es)
José Rojo Montijano (jrojo.eps@ceu.es)
Anselmo Romero Limón (arlimon@ceu.es)
Alfredo Sánchez Alberca (asalber@ceu.es)
Susana Victoria Rodríguez (victoria.eps@ceu.es)

Department of Applied Math and Statistics
CEU San Pablo

September 2016

CEU Universidad San Pablo

Applied Biostatistics with R
Alfredo Sánchez Alberca (asalber@ceu.es)

Contents

Practice 1

Introduction to R and RKWard

1 Introduction

The big computing power achieved by computers has made them powerful tools essentials for all those disciplines, such as Statistics, that require managing large volumes of data. Nowadays, nobody addresses a serious statistical study without the help of data analysis software.

R is a powerful programming language that includes a lot of functions for representing and analyzing data. It was developed by Robert Gentleman and Ross Ihaka at the university of Auckland in New Zealand, although now is maintained by a large scientific community all around the world.

The advantages of R with respect to other data analysis software like SPSS, SAS, Matlab, Minitqab or Excel, are multiple:

- It's open source software, what implies that it's free. It can be downloaded from the web `http://www.r-project.org/`.
- It's multi-platform . There are versions for the main platforms: Windows, Macintosh, Linux, etc.
- It's supported by a huge scientific community that uses the software as an standard for data analysis.
- It has thousands of packages for performing any type of data analysis and graphics representations, from the most common to the most innovative and cutting edge statistical procedures that are not included in other software. The packages are organized and documented in the CRAN (Comprehensive R Archive Network) repository, from where they can be downloaded for free. In Spain there is a mirror of this repository in the web `http://cran.es.r-project.org/`.
- It's programmable, what means that the user can create easily their own functions or packages for specific data analysis.
- There are a lot of books, manuals and tutorials for free, that allow the learning even for advances uses in all kind of disciplines like Mathematics, Physics, Chemistry, Biology, Psychology, Medicine, etc.

By default the working environment of R is the command line, what means that the computations are done with commands manually typed in a text window. However, there exists some graphical user interfaces (GUI) that ease its use, particularly for newbies. The GUI that we are going to use along these practices is *RKWard*, developed by Thomas Friedrichsmeier, and the package rk.Teaching developed by Alfredo Sánchez Alberca in the CEU San Pablo University and specifically designed for teaching Statistics.

The main goal of this chapter is to introduce the student to the use of RKWard and rk.Teaching, showing him the basic operations to enter and manipulate data.

2 Installation

2.1 Installation of R

Linux In the Debian distribution and derivatives (Ubuntu, Kubuntu, etc.) is as easy as typing the following command in the command prompt

```
> sudo apt-get install r-base-html r-cran-rcmdr r-cran-rodbc r-doc-html r
-recommended
```

Windows Download from http://cran.es.r-project.org/bin/windows/base/release.htm the installation program, execute it and follow the instructions.

2.2 Intallation of RKWard and rk.Teaching

The GUI RKWard can be downloaded from the web http://rkward.sourceforge.net/ where there are instructions for the different platforms.

For Windows it is recommended to install the full package that contains R, RKWard and rk.Teaching available in http://aprendeconalf.es/rkteaching/

By default, the installation of R includes the basic packages for the most common operations or analysis. However, for other analysis is required to install additional packages, like for instance the package rkTeaching, that includes menus and dialogs for most of the statistical procedures teached in these practices.

To install the package rk.Teching, first you have to download it from the web http://asalber.github.io/rkTeaching/, then start R or RKWard and enter the following command in the command console.

```
> setwd("path_to_download")
> install.packages("rk.Teaching",repos=NULL,dep=True)
```

For packages hosted in the CRAN, it's possible to install them from RKWard selecting the menu Settings > Manage R packages and plugins. In the dialog displayed, select the Install/Update/Remove R packages tab, then search or select the desired package and click the button OK. After installing a package, to use it in a work session it must be loaded. For that, in the same menu, select the Load/Unload R packages tab, then select the package, click the button Load and finally the button OK.

3 Solved exercises

1. Create a data set with the data in the sample below and save it with name colesterol.rda

Name	Gender	Weight	Height	Cholesterol
José Luis Martínez Izquierdo	H	85	179	182
Rosa Díaz Díaz	F	65	173	232
Javier García Sánchez	M	71	181	191
Carmen López Pinzón	F	65	170	200
Marisa López Collado	F	51	158	148
Antonio Ruiz Cruz	M	66	174	249

To create a data set:

(a) Select the menu File › New › Dataset.

(b) In the dialog displayed enter the name cholesterol and click the button OK.

(c) In the data editor window, define a variable in every column giving a name for the variable in the Name row, a type (Numeric, Factor, String or Logical) in the Type row, and in the case of a factor, defining its levels in the Levels row.

(d) After define the variable, enter the data of every variable in the sample in the corresponding column.

To save the data set:

(a) Select the menu Workspace › Save Workspace.

(b) In the dialog displayed give the name cholesterol.rda to the file, select the folder where to save it and click the button Save.

2. Define a new variable Age with the ages of the individuals in the sample, between the variables Name and Gender.

Name	Age
José Luis Martínez Izquierdo	18
Rosa Díaz Díaz	32
Javier García Sánchez	24
Carmen López Pinzón	35
Marisa López Collado	46
Antonio Ruiz Cruz	68

(a) Click the left tab Workspace.

(b) In the workspace window double-click the data set cholesterol to edit it.

(c) In the data editor window, right-click the column header of the variable Gender and select Insert new variable left.

(d) In the new empty column enter the name and type of the variable Age, and enter the age of every individual.

3. Insert the following data of a new individual:

Name: Cristóbal Campos Ruiz.
Age: 44 years.
Gender: Male.
Weight: 70 Kg.
Height: 178 cm.
Cholesterol: 220 mg/dl.

(a) Insert the data of the new individual in the first empty row.

4. Create a new variable with the body mass index of every individual using the formula

$$\text{bmi} = \frac{\text{Weight (in Kg)}}{\text{Height (in m)}^2}$$

(a) Select the menu Teaching › Data › Compute variable.
(b) In the dialog displayed enter the formula to compute the body mass index in the field Variable computation.
(c) In the field Save as click the button Change.
(d) In the dialog displayed select as a parent object the data set cholesterol and click the button OK.
(e) Enter the name bmi for the new variable and click the button Submit.

5. Recode the body mass index variable into a new categorical variable Obesity according to the following rules:

Less than de 18.5	→	Low weight
From 18.5 to 24.5	→	Healthy
From 24.5 to 30	→	Overweight
Greater than 30	→	Obese

(a) Select the el menu Teaching › Data › Variable recoding.
(b) In the dialog displayed insert the variable bmi in the field Variable to recode.
(c) Enter the recoding rules below in the field Recoding rules:

```
lo:18.5 = 1
18.5:24.5 = 2
24.5:30 = 3
30:hi = 4
```

(d) In the field Save as click the button Change.
(e) In the dialog displayed select as a parent object the data set cholesterol and click the button OK.
(f) Enter the name Obesity for the new variable and click the button Submit.
(g) In the data edition window, enter the levels for the Obesity factor, setting the label "Low weight" for the first category, "Healthy" for the second one, "Overweight" for the third and "Obese" for the fourth.

6. Filter the data set to get a new data set with the data of males.

(a) Select the menu Teaching › Data › Data filtering.
(b) In the dialog displayed insert the data set cholesterol in the field Data set.
(c) Insert the expression `gender=="M"` in the field Selection condition.
(d) Enter the name cholesterol.males for the new data set and click the button Submit.

4 Proposed exercises

1. The data set neonates of the package rk.Teaching, contains information about a sample of 320 newborns that meet the normal gestation time in a hospital during one year. Do the following operations:

    (a) Load the data set.

1. Click the Workspace tab and double-click the rk.Teaching package to unfold the data sets that it contains.
2. Right-click the data set nenonates and select the menu Copy to .GlobalEnv to copy the data set to the working environment.

(b) Compute the variable APGAR.average as the mean of the variables APGAR1 and APGAR5.

(c) Recode the variable weight into the factor weight.category with two categories corresponding to weights less than and greater than 2.5 Kg.

(d) Recode the variable APGAR1 into the factor APGAR.state with three categories: depressed (APGAR$\leq$ 3), moderately depressed (3 <APGAR$\leq$ 6) and normal (APGAR> 6).

(e) Filter the data set to get a new data set with the neonates of non-smoking mothers with an APGAR score at 1 minute less than or equal to 3. How many neonates are there?

Practice 2

Frequency distributions and charts

1 Solved exercises

1. The number of children in a sample of 25 families is

$$1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2, 0, 2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2.$$

Do the following operations:

(a) Create a data set with the variable children and enter the data.

(b) Create the frequency table.

> 1. Select the menu Teaching > Frequency distribution > Frequency table .
> 2. In the dialog displayed, select the variable children in the field Variable to tabulate and click the button Submit.

(c) Create the absolute frequency bar chart.

> 1. Select the menu Teaching > Charts > Bar chart.
> 2. In the dialog displayed, select the variable children in the field Variable and click the button Submit.

(d) Create also the relative frequency, cumulative absolute frequency and cumulative relative frequency bar charts, with their respective polygons.

> Follow the steps above checking, in the Bar options tab, the box Relative frequencies for the relative frequencies bar chart, the box Cumulative frequencies for the cumulative absolute bar chart, and both of them for the cumulative relative frequency bar chart. Check the box Polygon, to plot the corresponding polygon.

2. The number of people treated in the emergency service of a hospital every day of November was

15	23	12	10	28	7	12	17	20	21	18	13	11	12	26
30	6	16	19	22	14	17	21	28	9	16	13	11	16	20

Do the following operations:

(a) Create a data set with the variable emergencies and enter the data.

(b) Create the box plot. Are there some outlier? In that case, remove the outliers and proceed with the next part.

1. Select the menu Teaching › charts › Box plot.
2. In the dialog displayed select the variable emergencies in the field Variables and click the button Submit.
3. In the output windows with the box plot identify the outliers.
4. In the data set edition tab, remove the rows with the outliers right-clicking the row header and selecting Delete this row.

(c) Create the frequency table grouping data into 5 classes.

1. Select the menu Teaching › Frequency distribution › Frequency table.
2. In the dialog displayed select the variable emergencies.
3. In the Classes tab check the box Grouping intervals, check the option Number of intervals, enter the desired number of intervals in the field Suggested intervals and click the button Submit.

(d) Create the absolute frequency histogram.

1. Select the menu Teaching › charts › Histogram.
2. In the dialog displayed select the variable emergencies in the field Variable.
3. In the Classes tab, check the box Grouping intervals, check the box Number of intervals, enter the desired number of intervals in the field Suggested intervals and click the button Submit.

(e) Create also the relative frequency, cumulative absolute frequency and cumulative relative frequency histograms, with their respective polygons.

Follow the steps above checking, in the Histogram options, the box Relative frequencies for the relative frequency histogram, the box Cumulative frequencies for the cumulative absolute frequency histogram, and both of them for the cumulative relative frequency histogram. Check the box Polygon to plot the corresponding polygon.

3. The blood type of a sample of 30 persons are:

A, B, B, A, AB, 0, 0, A, B, B, A, A, A, A, AB,
A, A, A, B, 0, B, B, B, A, A, A, 0, A, AB, 0.

Do the following operations:

(a) Create a data set with the variable blood.type and enter the data.

(b) Create the frequency table.

1. Select the menu Teaching › Frequency distribution › Frequency table.
2. In the dialog displayed, select the variable blood.type in the field Variable to tabulate and click the button Submit.

(c) Create the pie chart.

1. Select the menu Teaching › charts › Pie chart.
2. In the dialog displayed, select the variable blood.type in the field Variables and click the button Submit.

4. The age and the marital status of a sample of 28 persons are:

Marital status	Age								
Single	31	45	35	65	21	38	62	22	31
Married	72	39	62	59	25	44	54		
Widow(er)	80	68	65	40	78	69	75		
Divorced	31	65	59	58	50				

Do the following operations:

(a) Create a data set with the variables marital.status and age and enter the data.

(b) Create the frequency table of the variable age for every marital status.

> 1. Select the menu Teaching › Frequency distribution › Frequency table.
> 2. In the dialog displayed, select the variable age in the field Variable to tabulate, check the box Tabulate by groups and select the variable marital.status in the field Grouping variable(s).
> 3. In the Classes tab, check the box Grouping intervals and click the button Submit.

(c) Create the box plots of age for every marital status. Are there outliers? Which group have more spread in ages?

> 1. Select the menu Teaching › charts › Box plot.
> 2. In the dialog displayed, select the variable age in the field Variable(s), check the box Plot by groups, select the variable marital.status in the field Grouping variable(s) and click the button Submit.

(d) Create the relative frequency histogram of age for every marital status. Compare the histograms.

> 1. Select the menu Teaching › charts › Histogram.
> 2. In the dialog displayed, select the variable age in the field Variable, check the box Plot by groups and select the variable marital.status in the field Grouping variable(s).
> 3. In the Classes tab, check the box Grouping intervals and click the button Submit.

2 Proposed exercises

1. The number of injuries suffered by the members of a soccer team in a league were

$$0, 1, 2, 1, 3, 0, 1, 0, 1, 2, 0, 1, 1, 1, 2, 0, 1, 3, 2, 1, 2, 1, 0, 1$$

Do the following operations:

(a) Construct the frequency table.

(b) Create the relative frequency and the cumulative relative frequency bar charts.

(c) Create the box plot.

2. The heights (in cm) of 30 students are

$$\begin{gathered}179, 173, 181, 170, 158, 174, 172, 166, 194, 185,\\ 162, 187, 198, 177, 178, 165, 154, 188, 166, 171,\\ 175, 182, 167, 169, 172, 186, 172, 176, 168, 187.\end{gathered}$$

Do the following operations:

(a) Create the absolute frequency histogram with classes of width 10 cm from 150 to 200 cm. Are there outliers?

3. The data set neonates of the package rk.Teaching, contains information about a sample of 320 newborns that meet the normal gestation time in a hospital during one year. Do the following operations:

   (a) Construct the frequency table of the APGAR score at 1 minute. If an score less than or equal to 3 indicates that the neonate is depressed, what percentage of neonates is depressed in the sample?

   (b) Construct the frequency table of the neonate weight, grouping into classes of width 0.5 Kg from 2 to 4.5 Kg. What intervals contains more neonates?

   (c) Compare the frequency distribution of the APGAR score at 1 minute for mothers less than 20 years old and for mothers greater than 20 years old. What group has more depressed neonates?

   (d) Compare the relative frequency distribution of the neonate weight according to whether the mother smoked or not during the pregnancy. If a weight under 2.5 Kg is considered a low weight, what group has a higher percentage of neonates with low weight?

   (e) Compute the prevalence of neonates with low weight for smoking and non-smoking mothers during the pregnancy.

   (f) Compute the relative risk of low weight of neonate when the mother smokes vs when then mother doesn't smoke during the pregnancy.

   (g) Create the bar chart of the APAGAR score at 1 minute. What is the more common score?

   (h) Construct the cumulative relative frequency bar chart of the APGAR score at 1 minute. Below what value is half of the neonates?

   (i) Compare the relative frequency distribution bar charts of the APGAR score at 1 minute according to whether the mother smoked or not during the pregnancy. What conclusion can be drawn?

   (j) Construct the histogram of the neonates weights with classes of width 0.5 Kg from 2 to 4.5 Kg. What class contains more neonates?

   (k) Compare the relative frequency histograms of the neonates weights, with classes of width 0.5 Kg from 2 to 4.5, Kg, according to whether the mother smoked or not during the pregnancy. What group has neonates with less weight?

   (l) Compare the relative frequency histograms of the neonates weights, with classes of width 0.5 Kg from 2 to 4.5, Kg, according to whether the mother smoked or not before the pregnancy. What conclusions can be drawn?

   (m) Construct the box plot of the neonates weights. What range of weights can be considered normal in the sample? Are there outliers in the sample?

   (n) Compare the box plots of the neonates weights according to whether the mother smoked or not during the pregnancy and whether the mother was less than 20 or greater than 20 years old. What group has more central spread? What group has neonates with less weight?

   (o) Compare the box plots of the APGAR scores at 1 minute and at 5 minutes. What variable has more central spread?

Practice 3

Sampling statistics

1 Solved exercises

1. The number of children in a sample of 25 families is

$$1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2, 0, 2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2.$$

Do the following operations:

(a) Create a data set with the variable children and enter the data.

(b) Compute the arithmetic mean, variance and standard deviation of the number of children. Interpret the statistics.

> 1. Select the menu Teaching > Descriptive statistics > Statistics.
> 2. In the dialog displayed insert the variable children in the field Variable.
> 3. In the Basic statistics tab check the boxes of Arithmetic mean, Variance and Standard deviation, and click the button Submit.

(c) Compute the quartiles, the range, the interquartile range, the third decile and the 68th percentile.

> 1. Select the menu Teaching > Descriptive statistics > Statistics.
> 2. In the dialog displayed insert the variable children in the field Variable.
> 3. In the Basic statistics tab check the boxes of Quartiles, Range, Interquartile range, enter the values 0.3 and 0.68 in the field Percentiles, and click the button Submit.

2. The number of people treated in the emergency service of a hospital every day of November was

15	23	12	10	28	7	12	17	20	21	18	13	11	12	26
30	6	16	19	22	14	17	21	28	9	16	13	11	16	20

Do the following operations:

(a) Create a data set with the variable emergencies and enter the data.

(b) Compute the arithmetic mean, variance, standard deviation and coefficient of variation of the number of emergencies. Interpret the statistics.

> 1. Select the menu Teaching > Descriptive statistics > Statistics.
> 2. In the dialog displayed insert the variable emergencies in the field Variable.
> 3. In the Basic statistics tab check the boxes of Arithmetic mean, Variance, Standard deviation and Coefficient of variation, and click the button Submit.

(c) Compute the coefficients of skewness and kurtosis and interpret the statistics.

1. Select the menu Teaching › Descriptive statistics › Statistics.
2. In the dialog displayed insert the variable emergencies in the field Variable.
3. In the Basic statistics tab check the boxes of Coefficient of skewness and Coefficient of kurtosis and click the button Submit.

3. In a group of 20 students the grades in Mathematics were

SS, AP, SS, AP, AP, NT, NT, AP, SB, SS
SB, SS, AP, AP, NT, AP, SS, NT, SS, NT

Do the following operations:

(a) Create a data set course with the variable grades and enter the data.

(b) Recode the grades into scores assigning 2.5 to SS, 6 to AP, 8 to NT and 9.5 to SB.

1. Select the menu Teaching › Data › Variable recoding.
2. In the dialog displayed insert the grades in the field Variable to recode.
3. Enter the following recoding rules in the field Recoding rules:

```
"SS"= 2.5
"AP"= 6
"NT"= 8
"SB"= 9.5
```

4. In the Save new variable click the button Change.
5. In the dialog displayed select as parent object the data set course and click the button Accept.
6. Enter the name score for the new variable, uncheck the box Convert in a factor and click the button Submit.

(c) Compute the median and the interquartile range.

1. Select the menu Teaching › Descriptive statistics › Statistics.
2. In the dialog displayed select the variable score in the field Variable.
3. In the Basic statistics tab check the boxes of Median and Interquartile range and click the button Submit.

4. The heights (in cm) of 30 students are

Females: 173, 158, 174, 166, 162, 177, 165, 154, 166, 182, 169, 172, 170, 168.
Males: 179, 181, 172, 194, 185, 187, 198, 178, 188, 171, 175, 167, 186, 172, 176, 187.

Do the following operations:

(a) Create a data set with the variables height and gender and enter the data.

(b) Compute the arithmetic mean, median, variance, standard deviation and quartiles according to the gender. Interpret the statistics.

1. Select the menu Teaching › Descriptive statistics › Statistics.
2. In the dialog displayed insert the variable height in the field Variable, check the box Statistics by groups and insert the variable gender in the field Grouping variable(s).
3. In the Basic statistics tab check the boxes of Arithmetic mean, Median, Variance, Standard deviation and Quartiles, and click the button Submit.

2 Proposed exercises

1. The number of injuries suffered by the members of a soccer team in a league were

   0, 1, 2, 1, 3, 0, 1, 0, 1, 2, 0, 1, 1, 1, 2, 0, 1, 3, 2, 1, 2, 1, 0, 1

   Do the following operations:

   (a) Compute la arithmetic mean, median, variance and standard deviation of the number of injuries and interpret them.

   (b) Compute the coefficients of skewness and kurtosis.

   (c) Compute the fourth and the eighth deciles and interpret them.

2. We want to compare the reliability of two blood pressure monitors, an arm monitor and a wrist monitor. For that purpose we have performed 8 repeated measures of the blood pressure of the same person with both moniors. The measurements (in mmHg) were:

   Arm monitor: 111, 109, 112, 111, 113, 113, 114, 111
   Wrist monitor: 115, 113, 117, 116, 112, 112, 117, 112

   Which monitor is more reliable?

3. The age and the marital status of a sample of 28 persons are:

Marital status	Age								
Single	31	45	35	65	21	38	62	22	31
Married	72	39	62	59	25	44	54		
Widow(er)	80	68	65	40	78	69	75		
Divorced	31	65	59	58	50				

   Do the following operations:

   (a) Compute the arithmetic mean and the standard deviation of the age according to the marital status and interpret them.

   (b) What group has the most representative mean?

4. A study wants to determine if there are relations between the blood pressure and the tobacco and drink. The values observed in a sample of 25 persons were:

Smokes	yes	no	yes	yes	yes	no	no	yes	no	yes	no	yes	no
Drinks	no	no	yes	yes	no	no	yes	yes	no	yes	no	yes	yes
Blood pressure	80	92	75	56	89	93	101	67	89	63	98	58	91

Smokes	yes	no	no	yes	no	no	no	yes	no	yes	no	yes
Drink	yes	no	yes	yes	no	no	yes	yes	yes	no	yes	no
Blood pressure	71	52	98	104	57	89	70	93	69	82	70	49

   (a) Compute the arithmetic mean, the standard deviation and the coefficients of skewness and kurtosis of the blood pressure for smokers and non-smokers, and interpret them.

   (b) Compute the same statistics for drinkers and non-drinkers. Interpret the statistics.

   (c) Compute the same statistics for smokers and drinkers, smokers and non-drinkers, non-smokers and drinkers, and non-smoker and non-drinkers. Interpret the statistics

Practice 4

Linear regression

1 Solved exercises

1. The values of two variables X and Y measured in a sample of 10 individuals are:

X	0	1	2	3	4	5	6	7	8	9
Y	2	5	8	11	14	17	20	23	26	29

Do the following operations:

(a) Create a data set with the variables X and Y and enter the data.

(b) Construct the scatter plot of X and Y.

> 1. Select the menu Teaching › Charts › Scatter plot.
> 2. In the dialog displayed, select the variable Y in the field Y variable, and the variable X in the field X variable, and click the button Submit.

According to the point cloud, what type of regression model explains better the relation between X and Y?

(c) Compute the regression line of Y on X.

> 1. Select the menu Teaching › Regression › Linear regression.
> 2. In the dialog displayed, insert the variable Y in the field Dependent variable and the variable X in the field Independent variable(s), and click the button Submit.

(d) Plot the regression line on the scatter plot.

> 1. Select the menu Teaching › Charts › Scatter plot.
> 2. In the dialog displayed, insert the variable Y in the field Y variable and the variable X in the field X variable.
> 3. In the Fitted line tab, check the box Linear and click the button Submit.

(e) Compute la regression line of X on Y and plot it on the scatter plot.

> Repeat the steps of the previous part but inserting the variable X in the field Dependent variable and the variable Y in the field Independent variable(s).

(f) How are the residuals? Comment the results.

2. A study pretends to determine the relation between the daily hours of study and the number of failed subjects in a course. The values of these variables in a sample of 30 students were:

Study hours	Failed subjects	Study hours	Failed subjects	Study hours	Failed subjects
3.5	1	2.2	2	1.3	4
0.6	5	3.3	0	3.1	0
2.8	1	1.7	3	2.3	2
2.5	3	1.1	3	3.2	2
2.6	1	2.0	3	0.9	4
3.9	0	3.5	0	1.7	2
1.5	3	2.1	2	0.2	5
0.7	3	1.8	2	2.9	1
3.6	1	1.1	4	1.0	3
3.7	1	0.7	4	2.3	2

Do the following operations:

(a) Create a data set with the variables study.hours and failed.subjects and enter the data of the sample.

(b) Construct the two-dimensional frequency table of the variables study.hours and failed.subjects.

> 1. Select the menu Teaching › Frequency distribution › Two-dimensional frequency table.
> 2. In the dialog displayed insert the variable study.hours in the field Variable to tabulate in rows and the variable failed.subjects in the field Variable to tabulate in columns.
> 3. In the Row classes tab, check the box Grouping intervals for the row variable, and click the button Submit.

(c) Compute la regression line of failed.subjects on study.hours and plot it.

> To compute the regression line:
>
> 1. Select the menu Teaching › Regression › Linear regression.
> 2. In the dialog displayed insert the variable failed.subjects in the field Dependent variable and the variable study.hours in the field Independent variable(s), check the box Save the model, enter the name linear.model.failed.subjects.on.study.hours for the regression model and click the button Submit.
>
> To plot he regression line:
>
> 1. Select the menu Teaching › Charts › Scatter plot.
> 2. In the dialog displayed insert the variable failed.subjects in the field Y variable and the variable study.hours in the field X variable.
> 3. In the Fitted line tab, check the box Linear and click the button Submit.

(d) What is the regression coefficient of the failed subjects on the daily hours of study? Interpret it.

> The regression coefficient is the slope of the regression line.

(e) The linear relation is stronger or weaker than in the previous exercise? Answer the question comparing the residuals in both linear models.

(f) Compute the linear coefficient of determination and the correlation coefficient. Is the linear model a good model to explain the relation between the failed subjects and the daily hours of study? What percentage of the variability of the failed subjects is explained by the linear model?

The coefficient of determination is showed as R^2 in output window, and the correlation coefficient is the square root.

(g) Use the linear model to predict the expected number of failed subjects for a student that studies 3 hours a day. Is this prediction reliable?

1. Select the menu Teaching › Regression › Predictions.
2. In the dialog displayed insert the model linear.model.failed.subjects.on.study.hours in the field Regression model, enter the value 3 in the in field Predictions for and click the button Submit.

(h) According to the linear model, how many hours of study are required at least to pass all the subjects?

To compute the regression line:

1. Select the menu Teaching › Regression › Linear regression.
2. In the dialog displayed insert the variable study.hours in the field Dependent variable and the variable failed.subjects in the field Independent variable(s), check the box Save the model, enter the name linear.model.study.hours.on.failed.subjects for the regression model and click the button Submit.

To make the prediction:

1. Select the menu Teaching › Regression › Predictions.
2. In the dialog displayed insert the model linear.model.study.hours.on.failed.subjects in the field Regression model, enter the value 0 in the in field Predictions for and click the button Submit.

3. To determine how an organism metabolizes the alcohol, an experiment was conducted where we measured the alcohol in blood every half an hour after drinking a liter of wine. The data of the experiment are below.

Time (min)	30	60	90	120	150	180	210
Alcohol (gr/l)	1.6	1.7	1.5	1.1	0.7	0.2	2.1

Do the following operations:

(a) Create a data set with the variables time and alcohol and enter the data of the sample.

(b) Compute the linear correlation coefficient of the alcohol and the time and interpret it. Is the linear model a good model to explain the metabolization of alcohol?

1. Select the menu Teaching › Regression › Linear regression.
2. In the dialog displayed insert the variable alcohol in the field Dependent variable, the variable time in the field Independent variable(s), and click the button Submit.

(c) Plot the regression line of alcohol on time. Are there some point with a big residual? In such a case, remove the point from the sample and compute again the linear correlation coefficient. Has the model improved?

1. Select the menu Teaching › Charts › Scatter plot.
2. In the dialog displayed insert the variable alcohol in the field Y variable, the variable time in the field X variable.
3. In the Fitted line tab click the box Linear and click the button Submit.

> It is observed that the point $(210, 2.1)$ has a huge residual compared to the others, what means that it's an outlier. To remove the outlier in the data edition windows, right-click the header row corresponding to the point and select Delete this row.

(d) If, according to the law, the maximum concentration of alcohol in blood to drive is 0.3 g/l, how much time must wait this person to drive after drinking a liter of wine? Is this prediction reliable?

> To compute the regression line:
>
> 1. Select the menu Teaching › Regression › Linear regression.
> 2. In the dialog displayed insert the variable time in the field Dependent variable and the variable alcohol in the field Independent variable(s).
> 3. Check the box Save the model, enter the name linear.model.time.on.alcohol for the linear model and click the button Submit.
>
> To make the prediction:
>
> 1. Select the menu Teaching › Regression › Predictions.
> 2. In the dialog displayed insert the model linear.model.time.on.alcohol in the field Regression model, enter the value 0.3 in the field Predictions for and click the button Submit.

4. The data set age.height of the package rk.Teaching contains the age and the height of 30 persons. Do the following operations:

(a) Load the data set age.height from the package rk.Teaching.

(b) Compute la regression line of the height on the age. Is the linear model a good model to explain the relation between the height and the age?

> 1. Select the menu Teaching › Regression › Linear regression.
> 2. In the dialog displayed insert the variable height in the field Dependent variable, the variable age in the field Independent variable(s), and click the button Submit.

(c) Create the scatter plot of the height on the age. Around which age changes the tendency?

> 1. Select the menu Teaching › Charts › Scatter plot.
> 2. In the dialog displayed insert the variable height in the field Y variable, the variable age in the field X variable and click the button Submit.

(d) Recode the variable age into the categorical variable age.group with two categories for younger and older than 20 years.

> 1. Select the menu Teaching › Data › Variable recoding.
> 2. In the dialog displayed insert the variable age in the field Variable to recode.
> 3. In the field Recoding rules enter the following rules:
>
> ```
> lo:20 = "younger"
> 20:hi = "older"
> ```
>
> 4. In the field Save new variable click the button Change.
> 5. In the dialog displayed select as parent object the data set age.height and click the button OK.
> 6. Enter the name age.group for the new variable and click the button Submit.

(e) Compute the regression line of the height on the age for every age group. In which group the linear model explains better the relation between the height and the age? Justify the answer.

1. Select the menu Teaching » Regression » Linear regression.
2. In the dialog displayed insert the variable height in the field Dependent variable and la variable age in the field Independent variable(s).
3. Check the box Regression by groups and insert the variable age.group in the field Grouping variable(s).
4. Check the box Save the model, enter the name linear.model.height.on.age for the linear model and click the button Submit.

(f) Plot the regression lines of the previous part.

1. Select the menu Teaching » Charts » Scatter plot.
2. In the dialog displayed insert the variable height in the field Y variable and the variable age in the field X variable.
3. Check the box Plot by groups and insert the variable age.group in the field Grouping variable(s).
4. In the Fitted line tab, check the box Linear and click the button Submit.

(g) According to the linear model, what is the expected height for a 14 years old person? And for a 38 years old person?

To predict the height of the 14 years old person:

1. Select the menu Teaching » Regression » Predictions.
2. In the dialog displayed insert the model linear.regression.height.on.age.younger in the field Regression model, enter the value 14 in the field Predictions for and click the button Submit.

To predict the height of the 38 years old person:

1. Select the menu Teaching » Regression » Predictions.
2. In the dialog displayed insert the model linear.regression.height.on.age.older in the field Regression model, enter the value 38 in the field Predictions for and click the button Submit.

2 Proposed exercises

1. A research study has been conducted to determine the loss of activity of a drug. The table below shows the results of the experiment.

Time (in years)	1	2	3	4	5
Activity (%)	96	84	70	58	52

Do the following operations:

(a) Compute the regression line of the drug activity on time.

(b) What percentage decreases the drug activity every year?

(c) How much time must pass for the drug to have an activity of 80%? When will be the activity null? Are these predictions equally reliable?

2. In an study about the effect of different doses of a medicament, 2 patients got 2 mg and took 5 days to cure, 4 patients got 2 mg and took 6 days to cure, 2 patients got 3 mg ant took 3 days to cure, 4 patients got 3 mg and took 5 days to cure, 1 patient got 3 mg and took 6 days to cure, 5 patients got 4 mg and took 3 days to cure and 2 patients got 4 mg and took 5 days to cure. Do the following operations:

   (a) Compute the regression line of the days to cure on the dose.

   (b) Compute the regression coefficient of the days to cure on the dose and interpret it.

   (c) Compute the correlation coefficient and interpret it.

   (d) Determine the expected time required to cure with a 5 mg dose. Is this prediction reliable? Justify the answer.

   (e) What dose must be applied to last 4 days to cure? Is this prediction reliable? Justify the answer.

3. The data set heights.weights.students of the package rk.Teaching, contains the height, the weight and the gender of a sample of students. Do the following operations:

   (a) Load the data set heights.weights.students from the package rk.Teaching.

   (b) Compute the regression line of weight on height and plot it.

   (c) Compute the regression lines of weight on height for males and females and plot them.

   (d) Compute the coefficients of determination for both models. Which model explains better the relation between weight and height, the males or the females one? Justify the answer.

   (e) What is the expected weight for a man 170 cm tall? And for a women of the same height?

4. The data set neonates of the package rk.Teaching, contains information about a sample of 320 newborns that meet the normal gestation time in a hospital during one year. Do the following operations:

   (a) Construct the two-dimensional frequency table of the APGAR score at 1 minute and whether the mother smoked or not during the pregnancy. What conclusions can you draw?

   (b) Construct the two-dimensional frequency table of te weight and the age of the mother. What conclusions can you draw?

   (c) Compute the regression line of the weight on the daily number of cigarettes smoked by the mother during the pregnancy. Is there a strong linear relation between the variables?

   (d) Plot the regression line of the previous part. Why the regression line doesn't fit well the point cloud?

   (e) Compute the regression line of the weight on the daily cigarettes smoked by the mother during the pregnancy in the group of smoking mothers. Is this regression model better or worse than the previous one? According to this model, how much decreases the weight of newborns for every daily cigarette smoked by the mother?

   (f) According to the previous linear model, what will be the expected weight of a nenonate with a mother that smokes 5 daily cigarettes during the pregnancy? And for a mother that smokes 30 daily cigarettes? Are these predictions reliable?

   (g) Are there the same linear relation between the weight and the daily cigarettes somoked by the mother for mothers younger than 20 years and mothers older than 20 years? What conclusions can you draw?

Practice 5

Non-linear regression

1 Solved exercises

1. The number of bacteria in a culture evolves with time according to the table below.

Time (hours)	0	1	2	3	4	5	6	7	8
Bacteria	25	28	47	65	86	121	190	290	362

Do the following operations:

(a) Create a data set with the variables time and bacteria and enter the data of the sample.

(b) Plot the scatter plot. According to the point cloud, what type of model explains better the bacteria evolution?

> 1. Select the menu Teaching › Charts › Scatter plot.
> 2. In the dialog displayed insert the variable bacteria in the field Y variable and the variable time in the field X variable, and click the button Submit.

(c) Compute the quadratic and exponential models of bacteria on time.

> To compute the quadratic model:
>
> 1. Select the menu Teaching › Regression › Non-linear regression.
> 2. In the dialog displayed insert the variable bacteria in the field Dependent variable and the variable time in the field Independent variable.
> 3. In the Regression models tab check the box Quadratic.
> 4. Check the box Save model, enter the name quadratic.model.bacteria.on.time and click the button Submit.
>
> To compute the exponential model:
>
> 1. Select the menu Teaching › Regression › Non-linear regression.
> 2. In the dialog displayed insert the variable bacteria in the field Dependent variable and the variable time in the field Independent variable.
> 3. In the Regression models tab check the box Exponential.
> 4. Check the box Save model, enter the name exponential.model.bacteria.on.time and click the button Submit.
>
> The best model is the one with a highest coefficient of determination.

(d) Plot the graph of the best model of the previous part.

1. Select the menu Teaching » Charts » Scatter plot.
2. In the dialog displayed insert the variable bacteria in the field Y variable and the variable time in the field X variable.
3. In the Fitted line tab check the box Exponencial and click the button Submit.

(e) According to the best model, how many bacteria there will be after 3 hours of the beginning of the culture? And after 10 hours? Are these predictions reliable?

1. Select the menu Teaching » Regression » Predictions.
2. In the dialog displayed insert the model exponential.model.bacteria.on.time in the field Regression model.
3. Enter the values 3.5, 10 in the field Predictions for and click the button Submit.
4. As it is an exponential model, the predictions are for the logarithm of bacteria. To get the prediction of bacteria you must apply the exponential function to the values obtained.

(f) Give a prediction as reliable as possible of the time required to have 100 bacteria in the culture.

To compute the logarithmic model:

1. Select the menu Teaching » Regression » Non-linear regression.
2. In the dialog displayed insert the variable time in the field Dependent variable and the variable bacteria in the field Independent variable.
3. In the Regression models tab check the box Logarithmic.
4. Check the box Save model, enter the name logarithmic.model.time.on.bacteria and click the button Submit.

To make the prediction:

1. Select the menu Teaching » Regression » Predictions.
2. In the dialog displayed insert the model logarithmic.model.time.on.bacteria in the field Regression model.
3. Enter the value 100 in the field Predictions for and click the button Submit.

2. The data set diet of the package rk.Teaching contains data of a study about a diet. For every individual it has been measured the number of days of diet, the weight loss and whether he or she does physical exercise regularly.

   Do the following operations:

   (a) Load the data set diet from the package rk.Teaching.

   (b) Plot the scatter plot. According to the point cloud, what type of model explains better the weight loss on the days of diet?

   1. Select the menu Teaching » Charts » Scatter plot.
   2. In the dialog displayed, select the variable weight.loss in the field Variable Y, la variable days in the field Variable X, and click the button Submit.

   (c) Compute the regression model that explains better the relation between the weight loss and the days of diet. Is it a good model for making predictions?

   1. Select the menu Teaching » Regression » Model comparison.
   2. In the dialog displayed insert the variable weight.loss in the field Dependent variable and the variable days in the field Independent variable.

3. In the Regression models tab check the boxes of the models to compare and click the button Submit.
4. The best model is the one with the greatest coefficient of determination.

(d) Plot the graph of the previous model.

1. Select the menu Teaching > Charts > Scatter plot.
2. In the dialog displayed insert the variable weight.loss in the field Y variable and the variable days in the field X variable.
3. In the tab Fitted line check the box of the corresponding model and click the button Submit.

(e) Compute the regression model that best explains the relation between the weight loss and the days of diet for the group of people who don't do physical exercise regularly.

To see what is the best regression model:

1. Select the menu Teaching > Regression > Model comparison.
2. In the dialog displayed insert the variable weight.loss in the field Dependent variable and the variable days in the field Independent variable.
3. Check the box Filter and enter the condition `exercise=="no"` in the field Selection condition.
4. In the Regression models tab check all the boxes and click the button Submit.
5. The best model is the one with the greatest coefficient of determination.

To compute the regression model:

1. Select the menu Teaching > Regression > Non-linear regression.
2. In the dialog displayed insert the variable weight.loss in the field Dependent variable and the variable days in the field Independent variable.
3. Check the box Filter and enter the condition `exercise=="no"` in the field Selection condition.
4. Check the box Save model, enter the name regression.weight.loss.on.days.no.exercise and click the button Submit.

(f) Compute the regression model that best explains the relation between the weight loss and the days of diet for the group of people who do physical exercise regularly.

To see what is the best regression mode:

1. Select the menu Teaching > Regression > Model comparison.
2. In the dialog displayed insert the variable weight.loss in the field Dependent variable and the variable days in the field Independent variable.
3. Check the box Filter and enter the condition `exercise=="yes"` in the field Selection condition.
4. In the Regression models tab check all the boxes and click the button Submit.
5. The best model is the one with the greatest coefficient of determination.

To compute the regression model:

1. Select the menu Teaching > Regression > Non-linear regression.
2. In the dialog displayed insert the variable weight.loss in the field Dependent variable and the variable days in the field Independent variable.
3. Check the box Filter and enter the condition `exercise=="yes"` in the field Selection condition.

4. Check the box Save model, enter the name regression.weight.loss.on.days.exercise and click the button Submit.

(g) Use the previous regression models to predict the weight loss after 40 and 300 days of diet for people who do physical exercise regularly and for people who don't. Are the predictions reliable?

Predictions for people who do exercise:

1. Select the menu Teaching › Regression › Predictions.
2. In the dialog displayed insert the model regression.weight.loss.on.days.exercise in the field Regression model.
3. Enter the values $40, 500$ in the field Predictions for and click the button Submit.

Predictions for people who don't do exercise:

1. Select the menu Teaching › Regression › Predictions.
2. In the dialog displayed insert the model regression.weight.loss.on.days.no.exercise in the field Regression model.
3. Enter the values $40, 500$ in the field Predictions for and click the button Submit.

2 Proposed exercises

1. The concentration of a drug in blood, in en mg/dl, depends on time according to the data below.

Time (hours)	2	3	4	5	6	7	8
Concentration	25	36	48	64	86	114	168

Do the following operations:

(a) According to the exponential modelSegún el modelo exponential, what will be the concentration of the drug in blood after hours? Is this prediction reliable?

(b) According to the logarithmic model, how much time must pass to have a concentration of 100 mg/dl of drug in blood?

Practice 6

Probability

1 Solved exercises

1. Construct the probability space of the following random experiments:

   (a) Draw a card from a Spanish deck of cards.

   > 1. Select the menu Teaching > Probability > Gambling > Cards > Probability space.
   > 2. In the dialog shown, enter 1 in the field Number of cards and click the button Submit.

   (b) Toss two coins.

   > 1. Select the menu Teaching > Probability > Gambling > Coins > Probability space.
   > 2. In the dialog shown, enter 2 in the field Number of coins and click the button Submit.

   (c) Roll two dice.

   > 1. Select the menu Teaching > Probability > Gambling > Dice > Probability space.
   > 2. In the dialog shown, enter 2 in the field Number of dice and click the button Submit.

   (d) Roll two dice and toss two coins.

   > 1. Select the menu Teaching > Probability > Combine independent probability spaces.
   > 2. In the dialog shown, select the data sets generated before corresponding to the probability spaces of rolling two dice and tossing two coins and click the button Submit.

2. Repeat the random experiment of tossing two coins 10 times, 100 times, 1000 times and 1000000 times and compute the relative frequency of every random event. Where does the frequencies tend to? Construct the probability space of the experiment and observe if it satisfies the law of large numbers, that is, that the frequency of each event tends to the probability of the event.

> To conduct the experiment:
>
> (a) Select the menu Teaching > Probability > Gambling > Coins > Tossing coins.
>
> (b) In the dialog shown, enter 2 in the field Number of coins, enter 10 in the field Number of repetitions, check the box Frequency distribution and click the button Submit.
>
> Repeat the previous steps but entering 100, 1000 and 1000000 respectively in the field Number of repetitions.
>
> To construct the probability space:

(a) Select the menu Teaching › Probability › Gambling › Coins › Probability space.
(b) In the dialog shown, enter 2 in the field Number of coins and click the button Submit.

3. In a cupboard there are three boxes of a medicine A, two boxes of medicine B and a box of medicine C. Construct the probability spaces of the following random experiments:

   (a) Pick three boxes randomly without replacement.

   1. Select the menu Teaching › Probability › Gambling › Urn › Probability space.
   2. In the dialog shown, check the box List objects, enter the values A,A,A,B,B,C in the field Object list, enter 3 in the field Number of extractions, and click the button Submit.

   (b) Pick three boxes randomly with replacement.

   Repeat the previous steps but checking the box With replacement.

4. An epidemiological investigation has been carried out in a population to determine the lifetime prevalence of three common diseases of childhood: chickenpox, measles and rubella. The observed frequencies appears in the table below.

Chickenpox	Measles	Rubella	Frequency
No	No	No	2654
No	No	Yes	1436
No	Yes	No	1682
No	Yes	Yes	668
Yes	No	No	1747
Yes	No	Yes	476
Yes	Yes	No	876
Yes	Yes	Yes	265

   (a) Create a data set chilhood.diseases with the variables chickenpox, measles, rubella and frequency and enter the data of the table.

   (b) Create the probability space of the population.

   1. Select the menu Teaching › Probability › Probability space.
   2. In the dialog shown enter the data set chilhood.diseases in the field Data set, check the box Define frequencies, enter the variable frequency in the Frequency, give the name chilhood.diseases.pe to the new data set containing the probability space and click the button Submit.

   (c) Compute the probability that a person of the population has suffered the chickenpox.

   1. Select the menu Teaching › Probability › Compute probability.
   2. In the dialog shown enter the data set chilhood.diseases.pe in the field Probability space, enter `chickenpox=="Yes"` in the field Event and click the button Submit.

   (d) Compute the probability that a person of the population has suffered the chickenpox or the measles.

   1. Select the menu Teaching › Probability › Compute probability.

2. In the dialog shown enter the data set chilhood.diseases.pe in the field Probability space, enter `chickenpox=="Yes" | measles=="Yes"` in the field Event and click the button Submit.

(e) Compute the probability that a person of the population has suffered the measles and the rubella.

1. Select the menu Teaching › Probability › Compute probability.
2. In the dialog shown enter the data set chilhood.diseases.pe in the field Probability space, enter `measles=="Yes" & rubella=="Yes"` in the field Event and click the button Submit.

(f) Compute the probability that a person of the population has suffered the chickenpox if he or she has suffered measles. Are independent the events of having suffered chickenpox and having suffered measles?

1. Select the menu Teaching › Probability › Compute probability.
2. In the dialog shown enter the data set chilhood.diseases.pe in the field Probability space, enter `chickenpox=="Yes"` in the field Event, check the box Conditional probability, enter `measles=="No"` in the field Condition and click the button Submit.

(g) Compute the probability that a person of the population has not suffered the rubella nor the measles if he or she has suffered the chickenpox.

1. Select the menu Teaching › Probability › Compute probability.
2. In the dialog shown enter the data set chilhood.diseases.pe in the field Probability space, enter `rubella=="No" & measles=="No"` in the field Event, check the box Conditional probability, enter `chickenpox=="Yes"` int the field Condition and click the button Submit.

5. A pregnancy test has been applied to a sample of women, getting the following results

Pregnancy	Test	Frequency
No	−	3876
No	+	47
Yes	−	12
Yes	+	131

(a) Create a data set pregnancy.test with the variables pregnancy, test, and frequency and enter the data of the table.

(b) Create a probability space from the sample.

1. Select the menu Teaching › Probability › Probability space.
2. In the dialog shown select the data set test.pregnancy, check the box Define frequencies, enter the variable frequency int the field Frequency, give the name pregnancy.test.pe to the new data set with the probability space and click the button Submit.

(c) Compute the prevalence of pregnancy.

1. Select the menu Teaching › Probability › Compute probability.
2. In the dialog shown enter the data set pregnancy.test.pe in the field Probability space, enter `pregnancy=="Yes"` in the field Event and click the button Submit.

(d) Compute the probability of having a positive outcome in the test.

1. Select the menu Teaching » Probability » Compute probability.
2. In the dialog shown enter the data set pregnancy.test.pe in the field Probability space, enter `test=="+"` in the field Event and click the button Submit.

(e) Compute the sensitivity of the test.

1. Select the menu Teaching » Probability » Compute probability.
2. In the dialog shown enter the data set pregnancy.test.pe in the field Probability space, enter `test=="+"` in the field Event, check the box Conditional probability, enter `pregnancy=="Yes"` in the field Condition and click the button Submit.

(f) Compute the specificity of the test.

1. Select the menu Teaching » Probability » Compute probability.
2. In the dialog shown enter the data set pregnancy.test.pe in the field Probability space, enter `test=="-"` in the field Event, check the box Conditional probability, enter `pregnancy=="No"` in the field Condition and click the button Submit.

(g) Compute the positive predictive value of the test. Is this test useful to detect a pregnancy?

1. Select the menu Teaching » Probability » Compute probability.
2. In the dialog shown enter the data set pregnancy.test.pe in the field Probability space, enter `pregnancy=="Yes"` in the field Event, check the box Conditional probability, enter `test=="+"` in the field Condition and click the button Submit.

(h) Compute the negative predictive value of the test. Is this test useful to rule out a pregnancy?

1. Select the menu Teaching » Probability » Compute probability.
2. In the dialog shown enter the data set pregnancy.test.pe in the field Probability space, enter `pregnancy=="No"` en el field Event, check the box Conditional probability, enter `test=="-"` in the field Condition and click the button Submit.

2 Proposed exercises

1. Create the sample space of the random experiment consisting on tossing a coin, rolling a die and drawing a card from a Spanish deck of cards.
2. To see the effectiveness of a vaccine against flu, a sample of 1000 persons was drawn from the a population. The table below summarize the number of persons that were or not vaccinated and that got or not the flu.

Vaccine	Flu	Frequency
No	No	418
No	Yes	312
Yes	No	233
Yes	Yes	37

   (a) Create a probability space from the sample.

(b) Compute the probability of having been vaccinated against the flu.
(c) Compute the prevalence of the flu.
(d) Compute the probability of having flu after having been vaccinated. Is the vaccine effective?

3. To see the effectiveness of a diagnostic test to diagnose ebola in a Central African country, the test was applied to a sample o persons. The outcome of the test was positive in 147 persons with ebola, but also in 28 persons without ebola. On the other hand, the outcome of the test was negative in 97465 persons without ebola, but also in 65 persons with ebola.

   (a) Create the probability space of the diagnostic test.
   (b) Compute the prevalence of ebola in the country.
   (c) Compute the probability of having a negative outcome in the test.
   (d) Compute the sensitivity and the specificity of the test.
   (e) Is more effective the test to detect the ebola or to rule out it?

Practice 7

Discrete Random Variables

1 Solved exercises

1. Let X be the variable that measures the number of heads got after tossing 10 coins, following a Binomial probability distribution model $B(10, 0.5)$.

   (a) Compute the probability distribution of X.

   > 1. Select the menu Teaching » Distributions » Discretes » Binomial » Probabilities.
   > 2. In the dialog shown, enter `0,1,2,3,4,5,6,7,8,9,10` in the field Values of the variable, enter `10` in the field Number of repetitions, `0.5` in the field Probability of success, and click the button Submit.

   (b) Plot the graph of the probability function of X.

   > 1. Select the menu Teaching » Distributions » Discretes » Binomial » Probability graph.
   > 2. In the dialog shown, enter `10` in the field Number of repetitions, `0.5` in the field Probability of success and click the button Submit.

   (c) Plot the graph of the distribution function of X.

   > 1. Select the menu Teaching » Distributions » Discretes » Binomial » Probability graph.
   > 2. In the dialog shown, enter `10` in the field Number of repetitions, `0.5` in the field Probability of success, check the box Distribution function and click the button Submit.

   (d) Compute the probability of getting 7 heads.

   > 1. Select the menu Teaching » Distributions » Discretes » Binomial » Probabilities.
   > 2. In the dialog shown, enter `7` in the field Values of the variable, enter `10` in the field Number of repetitions, `0.5` in the field Probability of success, and click the button Submit.

   (e) Compute the probability of getting less than 4 heads.

   > 1. Select the menu Teaching » Distributions » Discretes » Binomial » Probabilities.
   > 2. In the dialog shown, enter `4` in the field Values of the variable, `10` in the field Number of repetitions, `0.5` in the field Probability of success, check the box Cumulative probabilities and click the button Submit.

   (f) Compute the probability of getting more than 5 heads.

1. Select the menu Teaching › Distributions › Discretes › Binomial › Probabilities.
2. In the dialog shown, enter `5` in the field Values of the variable, `10` in the field Number of repetitions, `0.5` in the field Probability of success, check the box Cumulative probabilities, check the box Upper in the field Accumulation tail and click the button Submit.

(g) Compute the probability of getting two or more heads and less than 9.

1. Select the menu Teaching › Distributions › Discretes › Binomial › Probabilities.
2. In the dialog shown, enter the values `1,8` in the field Values of the variable, `10` in the field Number of repetitions, `0.5` in the field Probability of success, check the box Cumulative probabilities and click the button Submit.

The probability $P(2 \leq X < 9)$ is the difference between the probabilities obtained $P(X < 9) = P(X \leq 8)$ and $P(X < 2) = P(X \leq 1)$.

2. The number of births in a city X follows a Poisson probability distribution model with mean 6 births a day.

(a) Plot the graph of the probability function of X.

1. Select the menu Teaching › Distributions › Discretes › Poisson › Probability graph.
2. In the dialog shown, enter the value `6` in the field Mean and click the button Submit.

(b) Plot the graph of the distribution function of X.

1. Select the menu Teaching › Distributions › Discretes › Poisson › Probability graph.
2. In the dialog shown, enter the value `6` in the field Mean, check the box Distribution function and click the button Submit.

(c) Compute the probability that there is 1 birth a random day.

1. Select the menu Teaching › Distributions › Discretes › Poisson › Probabilities.
2. In the dialog shown, enter `1` in the field Values of the variable, enter `6` in the field Mean, and click the button Submit.

(d) Compute the probability that there are less than 6 births a random day.

1. Select the menu Teaching › Distributions › Discretes › Poisson › Probabilities.
2. In the dialog shown, enter `5` in the field Values of the variable, enter `6` in the field Mean, check the box Cumulative probabilities and click the button Submit.

(e) Compute the probability that there are 4 or more births a random day.

1. Select the menu Teaching › Distributions › Discretes › Poisson › Probabilities.
2. In the dialog shown, enter `3` in the field Values of the variable, enter `6` in the field Mean, check the box Cumulative probabilities, select the option Upper in the field Accumulation tail and click the button Submit.

(f) Compute the probability that there are between 4 and 8 births, both included, a random day.

1. Select the menu Teaching » Distributions » Discretes » Poisson » Probabilities.
2. In the dialog shown, enter `3,8` in the field Values of the variable, enter `6` in the field Mean, check the box Cumulative probabilities and click the button Submit.

The probability $P(4 \leq X \leq 8)$ is the difference between the probabilities obtained $P(X \leq 8)$ and $P(X < 4) = P(X \leq 3)$.

(g) Compute the probability that there are more than 30 and less than 40 births in a week.

1. Select the menu Teaching » Distributions » Discretes » Poisson » Probabilities.
2. In the dialog shown, enter `30,39` in the field Values of the variable, enter `42` in the field Mean, check the box Cumulative probabilities and click the button Submit.

The probability $P(30 < X < 40)$ is the difference between the probabilities obtained $P(X < 40) = P(X \leq 39)$ and $P(X \leq 30)$.

3. The law of rare events asserts that the Binomial probability distribution model $B(n, p)$, tends to the Poisson probability distribution model $P(np)$ when n tends to ∞ and p tends to 0. In particular, the Poisson model is a good approximation of the Binomial model for $n \geq 30$ and $p \leq 0.1$. To check this law,

(a) Compute the probability distribution of the Binomial model $B(30, 0.1)$.

1. Select the menu Teaching » Distributions » Discretes » Binomial » Probabilities.
2. In the dialog shown, enter `0,1,2,3,4,5,6,7,8,9,10` in the field Values of the variable, enter `30` in the field Number of repetitions, `0.1` in the field Probability of success and click the button Submit.

(b) Compute the probability distribution of the Poisson model $P(3)$ and compare it with the Binomial distribution $B(30, 0.1)$.

1. Select the menu Teaching » Distributions » Discretes » Poisson » Probabilities.
2. In the dialog shown, enter `0,1,2,3,4,5,6,7,8,9,10` in the field Values of the variable, enter `3` in the field Mean and click the button Submit.

(c) Compute the probability distribution of the Binomial model $B(100, 0.03)$ and compare it with the Poisson distribution $P(3)$. Are these distributions models more similar than the previous ones?

1. Select the menu Teaching » Distributions » Discretes » Binomial » Probabilities.
2. In the dialog shown, enter `0,1,2,3,4,5,6,7,8,9,10` in the field Values of the variable, enter `100` in the field Number of repetitions, `0.03` in the field Probability of success and click the button Submit.

(d) Plot the graphs of the probability functions of the previous models. Increase number of repetitions and decrease the probability of success in the Binomial model and observe how the probabilities of the Binomial and the Poisson models are more similar.

1. Select the menu Teaching » Simulations » Law of rare events.
2. In the dialog shown, enter `30` in the field n and enter `0.1` in the field p.
3. Then increase the value of n up to `100` and decrease the value of p down to `0.03`.

2 Proposed exercises

1. What is the probability of getting between 40 and 60 heads, both included, after tossing 100 coins?
2. The chance of being cured with a treatment is 0.85. If we apply the treatment to 6 patients,
   (a) Plot the graph of the probability function of the number of patients cured.
   (b) What is the probability that half of them are cured?
   (c) What is the probability that a least 4 of them are cured?
3. The probability of having an adverse reaction to a vaccine is 0.001. If 2000 persons are vaccinated, what is the probability of having some adverse reaction?
4. The average number of calls per minute that arrive to a telephone switchboard is 120.
   (a) What is the probability of receiving less than 4 calls in 2 seconds?
   (b) What is the probability of receiving at least 3 calls in 3 seconds?

Practice 8

Continuous Random Variables

1 Solved exercises

1. Suppose that a bus passes by a bus stop every 15 minutes and that a person can arrive at the bus stop at any moment with the same likelihood. Then, the variable that measures the waiting time for the bus follows an Uniform probability distribution model $U(0, 15)$, since any waiting time between 0 and 15 minutes has the same likelihood of happening.

   (a) Plot the graph of the density function of the waiting time.

   > 1. Select the menu Teaching › Distributions › Continuous › Uniform › Probability graph.
   > 2. In the dialog shown, enter `0` in the field Minimum, enter `15` in the field Maximum and click the button Submit.

   (b) Plot the graph of the distribution function of the waiting time.

   > 1. Select the menu Teaching › Distributions › Continuous › Uniform › Probability graph.
   > 2. In the dialog shown, enter `0` in the field Minimum, enter `15` in the field Maximum, check the box Distribution function and click the button Submit.

   (c) Compute the probability of waiting for the bus less than 5 minutes.

   > 1. Select the menu Teaching › Distributions › Continuous › Uniform › Probabilities.
   > 2. In the dialog shown, enter `5` in the field Values of the variable, enter `0` in the field Minimum, enter `15` in the field Maximum and click the button Submit.

   (d) Compute the probability of waiting for the bus more than 12 minutes.

   > 1. Select the menu Teaching › Distributions › Continuous › Uniform › Probabilities.
   > 2. In the dialog shown, enter `12` in the field Values of the variable, enter `0` in the field Minimum, enter `15` in the field Maximum, check the box Upper in the field Accumulation tail and click the button Submit.

   (e) Compute the probability of waiting for the bus between 5 and 10 minutes.

   > 1. Select the menu Teaching › Distributions › Continuous › Uniform › Probabilities.
   > 2. In the dialog shown, enter `5, 10` in the field Values of the variable, enter `0` in the field Minimum, enter `15` in the field Maximum and click the button Submit.
   >
   > La probability $P(5 \leq X \leq 10)$ is the difference between the probabilities obtained $P(X \leq 10) - P(X \leq 5)$.

(f) Compute the time such that half of the times the person have to wait for the bus less than that time.

1. Select the menu Teaching > Distributions > Continuous > Uniform > Quantiles.
2. In the dialog shown, enter 0.5 in the field Cumulative probabilities, enter 0 in the field Minimum, enter 15 in the field Maximum and click the button Submit.

(g) Compute the time such that 10% of the times the person have to wait for the bus more than that time.

1. Select the menu Teaching > Distributions > Continuous > Uniform > Quantiles.
2. In the dialog shown, enter 0.1 in the field Cumulative probabilities, enter 0 in the field Minimum, enter 15 in the field Maximum, check the box Upper in the field Accumulation tail and click the button Submit.

2. The random variable following a Normal probability distribution model with mean 0 and standard deviation 1, $Z \sim N(0,1)$, it is known as the Standard Normal.

(a) Plot the graph of the density function of Z.

1. Select the menu Teaching > Distributions > Continuous > Normal > Probability graph.
2. In the dialog shown, enter 0 in the field Mean, enter 1 in the field Standard deviation and click the button Submit.

(b) How does affect the mean and the standard deviation to the shape of the Gauss bell?

1. Select the menu Teaching > Distributions > Continuous > Normal > Probability graph.
2. In the dialog shown, check the box Preview.
3. Change the value of the mean and observe how changes the shape of the Gauss bell.
4. Then change the value of the standard deviations and observe how changes the shape of the Gauss bell.

(c) Plot the graph of the distribution function of Z.

1. Select the menu Teaching > Distributions > Continuous > Normal > Probability graph.
2. In the dialog shown, enter 0 in the field Mean, enter 1 in the field Standard deviation, check the box Distribution function and click the button Submit.

(d) Compute the probability $P(Z < -1)$.

1. Select the menu Teaching > Distributions > Continuous > Normal > Probabilities.
2. In the dialog shown, enter -1 in the field Values of the variable, enter 0 in the field Mean, enter 1 in the field Standard deviation, and click the button Submit.

(e) Compute the probability $P(Z > 1)$.

1. Select the menu Teaching > Distributions > Continuous > Normal > Probabilities.
2. In the dialog shown, enter 1 in the field Values of the variable, enter 0 in the field Mean, enter 1 in the field Standard deviation, check the box Upper in the field Accumulation tail and click the button Submit.

(f) Compute the probability that Z takes a value between the mean minus the standard deviation and the mean plus the standard deviation, that is, $P(-1 \leq Z \leq 1)$.

1. Select the menu Teaching » Distributions » Continuous » Normal » Probabilities.
2. In the dialog shown, enter `-1, 1` in the field Values of the variable, enter `0` in the field Mean, enter `1` in the field Standard deviation, and click the button Submit.

The probability $P(-1 \leq Z \leq 1)$ is the difference between the probabilities obtained $P(Z \leq 1) - P(Z \leq -1)$.

(g) Compute the probability that Z takes a value between the mean minus two times the standard deviation and the mean plus two times the standard deviation, that is, $P(-2 \leq Z \leq 2)$.

1. Select the menu Teaching » Distributions » Continuous » Normal » Probabilities.
2. In the dialog shown, enter `-2, 2` in the field Values of the variable, enter `0` in the field Mean, enter `1` in the field Standard deviation, and click the button Submit.

The probability $P(-2 \leq Z \leq 2)$ is the difference between the probabilities obtained $P(Z \leq 2) - P(Z \leq -2)$.

(h) Compute the probability that Z takes a value between the mean minus three times the standard deviation and the mean plus three times the standard deviation, that is, $P(-3 \leq Z \leq 3)$.

1. Select the menu Teaching » Distributions » Continuous » Normal » Probabilities.
2. In the dialog shown, enter `-3, 3` in the field Values of the variable, enter `0` in the field Mean, enter `1` in the field Standard deviation, and click the button Submit.

The probability $P(-3 \leq Z \leq 3)$ is the difference between the probabilities obtained $P(Z \leq 3) - P(Z \leq -3)$.

(i) Compute the quartiles.

1. Select the menu Teaching » Distributions » Continuous » Normal » Quantiles.
2. In the dialog shown, enter `0.25, 0.5, 0.75` in the field Cumulative probabilities, enter `0` in the field Mean, enter `1` in the field Standard deviation and click the button Submit.

(j) Compute the value of the standard normal with a lower probability tail of 0.95.

1. Select the menu Teaching » Distributions » Continuous » Normal » Quantiles.
2. In the dialog shown, enter `0.95` in the field Cumulative probabilities, enter `0` in the field Mean, enter `1` in the field Standard deviation and click the button Submit.

(k) Compute the value of the standard normal with an upper probability tail of 0.025.

1. Select the menu Teaching » Distributions » Continuous » Normal » Quantiles.
2. In the dialog shown, enter `0.025` in the field Cumulative probabilities, enter `0` in the field Mean, enter `1` in the field Standard deviation, check the box Upper in the field Accumulation tail and click the button Submit.

3. If $X_1, \ldots, X_n$ are n independent standard normal variables, then the variable $X = X_1^2 + \ldots + X_n^2$ follows a probability distribution model Chi-square with n degrees of freedom $\chi^2(n)$. Let X be a variable following a Chi-square probability distribution model with 6 degrees of freedom, $\chi^2(6)$.

   (a) Plot the graph of the density function of X.

1. Select the menu Teaching » Distributions » Continuous » Chi-square » Probability graph.
2. In the dialog shown, enter 6 in the field Degrees of freedom and click the button Submit.

(b) Compute the probability $P(X < 6)$.

1. Select the menu Teaching » Distributions » Continuous » Chi-square » Probabilities.
2. In the dialog shown, enter 6 in the field Values of the variable, enter 6 in the field Degrees of freedom and click the button Submit.

(c) Compute the 5th percentile.

1. Select the menu Teaching » Distributions » Continuous » Chi-square » Quantiles.
2. In the dialog shown, enter `0.05` in the field Cumulative probabilities, enter 6 in the field Degrees of freedom and click the button Submit.

(d) Compute el value with an upper probability tail of 0.1.

1. Select the menu Teaching » Distributions » Continuous » Chi-square » Quantiles.
2. In the dialog shown, enter `0.1` in the field Cumulative probabilities, enter 6 in the field Degrees of freedom, check the box Upper in the field Accumulation tail and click the button Submit.

4. If Y is a Chi-square variable with n degrees of freedom and Z a standard normal variable independents, then the variable $X = \frac{Z}{\sqrt{Y/n}}$ follows a Student's t probability distribution model with n degrees of freedom, $T(n)$. Let X be a variable following a Student's t probability distribution model with 8 degrees of freedom, $T(8)$.

(a) Plot the graph of the density function of X and compare it with the standard normal one.

1. Select the menu Teaching » Distributions » Continuous » Student's t » Probability graph.
2. In the dialog shown, enter 8 in the field Degrees of freedom and click the button Submit.

(b) Compute the 8th percentile.

1. Select the menu Teaching » Distributions » Continuous » Student's t » Quantiles.
2. In the dialog shown, enter `0.08` in the field Cumulative probabilities, enter 8 in the field Degrees of freedom and click the button Submit.

(c) Compute el value such that 5% of the population is above that value.

1. Select the menu Teaching » Distributions » Continuous » Student's t » Quantiles.
2. In the dialog shown, enter `0.05` in the field Cumulative probabilities, enter 8 in the field Degrees of freedom, check the box Upper in the field Accumulation tail and click the button Submit.

5. If Y_1 and Y_2 are two independent Chi-square variables with n and m degrees of freedom respectively, then the variable

$$X = \frac{Y_1/n}{Y_2/m}$$

follows a Fishers' F probability distribution model with n and m degrees of freedom, $F(n, m)$. Let X be a variable following a Fisher's F probability distribution model with 10 and 20 degrees of freedom, $F(10, 20)$.

(a) Plot the graph of the density function of X.

1. Select the menu Teaching » Distributions » Continuous » Fishers' F » Probability graph.
2. In the dialog shown, enter 10 in the field Numerator degrees of freedom, enter 20 in the field Denominator degrees of freedom and click the button Submit.

(b) Compute the probability $P(X > 1)$.

1. Select the menu Teaching » Distributions » Continuous » Fishers' F » Probabilities.
2. In the dialog shown, enter 1 in the field Values of the variable, enter 10 in the field Numerator degrees of freedom, enter 20 in the field Denominator degrees of freedom, check the box Upper in the field Accumulation tail and click the button Submit.

(c) Compute the interquartile range.

1. Select the menu Teaching » Distributions » Continuous » Fishers' F » Quantiles.
2. In the dialog shown, enter `0.25, 0.75` in the field Cumulative probabilities, enter 10 in the field Numerator degrees of freedom, enter 20 in the field Denominator degrees of freedom and click the button Submit.

The interquartile range is the difference between the third and the first quartiles.

2 Proposed exercises

1. It is known that the glucose level in blood of diabetic persons follows a normal distribution model with mean 106 mg/100 ml and standard deviation 8 mg/100 ml.
   (a) Calculate the probability of a random diabetic person having a glucose level less than 120 mg/100 ml.
   (b) What percentage of persons have a glucose level between 90 and 120 mg/100 ml?
   (c) Calculate and interpret the first quartile of the glucose level.
2. It is known that the cholesterol level in males 30 years old follows a normal distribution with mean 220 mg/dl and standard deviation 30 mg/dl. If there are 20000 males 30 years old in the population,
   (a) how many of them have a cholesterol level between 210 and 240 mg/dl?
   (b) If a cholesterol level greater than 250 mg/dl can provoke a thrombosis, how many of them are in risk of thrombosis?
   (c) Calculate the cholesterol level above which 20% of the males are?

Practice 9

Confidence intervals for one population

1 Solved exercises

1. The active ingredient concentration of a random sample of 10 drug containers drawn from a batch are (in mg/mm^3)

$$17.6 \quad 19.2 \quad 21.3 \quad 15.1 \quad 17.6 \quad 18.9 \quad 16.2 \quad 18.3 \quad 19.0 \quad 16.4$$

Do the following operations:

(a) Create a data set with the variable concentration and enter the data of the sample.

(b) Compute the confidence interval for the mean of the active ingredient concentration with a 95% confidence level (significance level $\alpha = 0.05$).

> 1. Select the menu Teaching > Parametric tests > Means > t-test for the mean of one population.
> 2. In the dialog displayed insert the variable concentration in the field Mean of and click the button Submit.

(c) Compute the confidence interval for the mean of the active ingredient concentration 99% confidence level (significance level $\alpha = 0.01$).

> 1. Select the menu Teaching > Parametric tests > Means > t-test for the mean of one population.
> 2. In the dialog displayed insert the variable concentration in the field Mean of·
> 3. In the Test options tab enter 0.99 in the Confidence level and click the button Submit.

(d) If we define the precision of the interval as the inverse of the its width, how changes the precision of an interval when we increase the confidence level? Why?

(e) What sample size is required to get an estimate of the mean of the active ingredient concentration with an error ± 0.5 mg/mm^3 and a 95% confidence level?

> 1. Select the menu Teaching > Descriptive statistics > Statistics.
> 2. In the dialog displayed insert the variable concentration in the field Mean of.
> 3. En the Basic statistics check the box Corrected standard deviation and click the button Submit.
> 4. Select the menu Teaching > Parametric tests > Means > Sample size to estimate one mean.
> 5. In the dialog displayed insert the corrected standard deviation in the field Standard deviation, enter 0.05 in the field Significance level, enter 0.5 in the field Error and click the button Submit.

(f) If the active ingredient concentration must be at least 16 mg/mm^3 in order to be effective, can we validate the batch? Justify the answer.

2. A diary company receive milk from two farms X and Y. To analyze the quality of milk, the milk fat have been measure for two samples of milk, one from each farm. The results are in the table below.

X		Y	
0.34	0.34	0.28	0.29
0.32	0.35	0.30	0.32
0.33	0.33	0.32	0.31
0.32	0.32	0.29	0.29
0.33	0.30	0.31	0.32
0.31	0.32	0.29	0.31
		0.33	0.32
		0.32	0.33

   (a) Create a data set with the variables fat and farm and enter the data of the sample.
   (b) Compute the 95% confidence interval for the mean of fat regardless the farm.

   1. Select the menu Teaching › Parametric tests › Means › t-test for the mean of one population.
   2. In the dialog displayed insert the variable fat in the field Mean of and click the button Submit.

   (c) Compute the 95% confidence intervals for the mean of fat for every farm.

   1. Select the menu Teaching › Parametric tests › Means › t-test for the mean of one population.
   2. In the dialog displayed select the variable fat in the field Mean of.
   3. Check the box Means by groups, insert the variable farm in the field Grouping variables(s) and click the button Submit.

   (d) Plot the 95% confidence intervals for the mean of fat for every farm.

   1. Select the menu Teaching › Charts › Means plot.
   2. In the dialog displayed select the variable fat in the field Mean(s) of.
   3. Check the box Plot by groups, insert the variable farm in the field Grouping variable(s) and click the button Submit.

   (e) Is there a significant difference between the milk fat means of the farms? Justify the answer.

3. In a survey performed by a university about the use of the library, a random sample of 34 students has been asked whether they go to the library at least once a week. The answers are shown below.

no	yes	no	no	no	yes	no	yes	yes	yes	yes	no	yes	no	yes	no	no
no	yes	yes	yes	no	no	yes	no	no	yes	yes	no	no	yes	no	yes	no

   (a) Create a data set with the variable answer and enter the data of the sample.
   (b) Compute the confidence interval for the proportion of students that uses the library at least once a week with a significance level 0.01, and interpret it.

   1. Select the menu Teaching › Parametric tests › Proportions › Test for one proportion.
   2. In the dialog displayed insert the variable answer in the field Variable and enter yes in the field Proportion of.
   3. In the Test options tab enter 0.99 in the field Confidence level and click the button Submit.

   (c) How is the precision of the confidence interval?
   (d) What sample size is required to get an estimate of the proportion of students that uses the library at least once a week with and error $\pm 1\%$ and a 95% confidence level?

1. Select the menu Teaching › Parametric tests › Proportions › Sample size to estimate one proportion.
2. In the dialog displayed enter the sample proportion in the field p, enter 0.05 in the field Significance level, enter 0.01 in the field Error and click the button Submit.

4. The Ministry of Health wants to compute a confidence interval for the proportion of people over 65 that with respiratory problems that have been vaccinated. In a random sample of 200 persons over 65 with respiratory problems, 154 were vaccinated.

   (a) Compute el 95% confidence interval for the proportion of people over 65 with respiratory problems vaccinated.

   1. Select the menu Teaching › Parametric tests › Proportions › Test for one proportion.
   2. In the dialog displayed check the box Manual entry of frequencies, enter 154 in the field Sample frequency, enter 200 in the field Sample size and click the button Submit.

   (b) If Ministry of Health goal si to achieve at least a 70% of people over 65 with respiratory problems vaccinated, can we say that the Ministry has achieved the goal? Justify the answer.

2 Proposed exercises

1. The level of cholesterol (in mg/dl) of a random sample of 8 persons of a population is

   196 212 188 206 203 210 201 198

   Compute the confidence intervals for the mean with significance levels 0.1, 0.05 and 0.01. Can we conclude that the mean of the level of cholesterol of the population is under 210 mg/dl?

2. To treat a neurological syndrome there are two techniques A and B. In a study a random sample of 60 persons were drawn. Technique A was applied to 25 of them and technique B to the others 35. 18 of the persons treated with A were cured, while 21 of the persons treated with B were cured. Compute the confidence interval for the proportion of persons that were cured with every technique. Which interval is more precise?

3. The data set neonates of the package rk.Teaching, contains information about a sample of 320 newborns that meet the normal gestation time in a hospital during one year. Do the following operations:

   (a) Compute the 99% confidence interval for the mean of the neonates weight.
   (b) Compute the confidence intervals for the APGAR score at 1 minute and for the APGAR score at 5 minutes and compare them. Is there a significant difference between the means of both scores?
   (c) Compute the confidence intervals for the percentage of neonates with weight less than or equal to 2.5 Kg for smoker and non-smoker mothers and compare them.

Practice 10

Confidence intervals for comparing two populations

1 Solved exercises

1. In order to see whether an advertising campaign has increased the sales of a drug, a sample of 8 pharmacies were drawn from a city. In each pharmacy the monthly sales of the drug before and after the campaign was recorded in the following table.

Before	147	163	121	205	132	190	176	147
After	150	171	132	208	141	184	182	145

(a) Create a data set with the variables before and after and enter the data.

(b) Compute the mean of monthly sales, before and after the campaign. Are the meas different? Have the campaign increased the sales of the drug?

> 1. Select the menu Teaching › Descriptive statistics › Statistics.
> 2. In the dialog displayed insert the variables before and after in the field Variables.
> 3. In the Basic statistics tab check the box Mean and click the button Submit.

(c) Compute the confidence intervals for the mean of the difference between the monthly sales after and before with confidence levels 0.05 and 0.01.

> 1. Select the menu Teaching › Parametric tests › Means › t-test for comparing the means of two paired populations.
> 2. In the dialog displayed insert the variable before in the field Compare mean of population and the variable after in the field With mean of population.
> 3. In the Test options enter 0.95 in the field Confidence level and click the button Submit.
> 4. Repeat the previous steps but entering 0.99 in the field Confidence level.

(d) Can we affirm that the advertising campaign has increased the drug sales significantly? Can we conclude the same if we change the sales after the campaign of the two last pharmacies putting 190 instead of 182 and 165 instead of 145? What happens to the confidence intervals?

> 1. Select the menu Teaching › Parametric tests › Means › t-test for comparing the means of two paired populations.
> 2. In the dialog displayed insert the variable before in the field Compare and the variable after in the field With.
> 3. In the Test options enter 0.95 in the field Confidence level and click the button Submit.
>
> We can conclude that there are significant differences between the means, with the confidence level set, if the confidence interval doesn't contain zero.

2. A diary company receive milk from two farms X and Y. To analyze the quality of milk, the milk fat have been measure for two samples of milk, one from each farm. The results are in the table below.

X		Y	
0.34	0.34	0.28	0.29
0.32	0.35	0.30	0.32
0.33	0.33	0.32	0.31
0.32	0.32	0.29	0.29
0.33	0.30	0.31	0.32
0.31	0.32	0.29	0.31
		0.33	0.32
		0.32	0.33

(a) Create a data set with the variables fat and farm and enter the data of the sample.

(b) Compute the confidence interval for the difference between the milk fat means of farms X and Y.

1. Select the menu Teaching › Parametric tests › Means › t-test for comparing the means of two independent populations.
2. In the dialog displayed insert the fat in the field Compare mean of and the variable farm in the field According to.
3. In the frame Populations to compare insert the value X in the field Compare and the value Y in the field With.
4. In the Test options enter 0.95 in the field Confidence level and click the button Submit.

(c) Can we conclude that the difference between the milk fat means of the farms is significant? Which farm has milk with more fat? How much more fat has the milk of X farm than the milk of Y farm?

We can conclude that there are significant differences between the means, with the confidence level set, if the confidence interval doesn't contain zero.

3. In a survey performed by a university about the use of the library, a random sample of 34 students has been asked whether they go to the library at least once a week. The answers and the gender of the students are shown below.

Answer	no	yes	no	no	no	yes	no	yes	yes	yes	yes	no	yes	no	yes	no	no
Gender	m	f	f	m	m	m	f	f	f	f	m	m	f	m	f	m	m
Answer	no	yes	yes	yes	no	no	yes	no	no	yes	yes	no	no	yes	no	yes	no
Gender	f	m	f	f	f	m	f	m	m	f	f	m	m	f	f	f	m

(a) Create a data set with the variables answer and gender.

(b) Compute the confidence interval for the difference between the proportions of females and males that use the library at least once a week.

1. Select the menu Teaching › Parametric tests › Proportions › Test for comparing two proportions.
2. In the dialog displayed insert the variable answer in the field Compare and the variable gender in the field According to.
3. Insert the value yes in the field Proportion of.
4. In the frame Populations to compare insert the value f in the field Compare population and the value m in the field With population.
5. In the Test options enter 0.95 in the field Confidence level and click the button Submit.

> We can conclude that there are significant differences between the proportions, with the confidence level set, if the confidence interval doesn't contain zero.

4. In a course there are two groups of students, one in the morning and the other in the afternoon. In the morning group 55 students of 80 passed, while in the afternoon group 32 students of 90 passed. Are there significant differences between the percentages of students that passed in the morning and in the afternoon? Can we conclude that the timetable is the cause of the differences? Justify the answer.

> (a) Select the menu Teaching › Parametric tests › Proportions › Test for comparing two proportions.
>
> (b) In the dialog displayed check the box Manual entry of frequencies, enter 55 in the field Sample frequency 1, enter 80 in the field Sample size 1, enter 32 in the field Sample frequency 2, enter 90 in the field Sample size 2 and click the button Submit.

2 Proposed exercises

1. In a study to determine the relation between the physical exercise and the level cholesterol in blood, a sample of 11 persons was drawn. The participants cholesterol level (in mg/dl) before and after doing a program of physical exercises is shown below.

Cholesterol before	182	232	191	200	148	249	276	213	241	280	262
Cholesterol after	198	210	194	220	138	220	219	161	210	213	226

   (a) Compute the 95% confidence interval for the mean of the difference of cholesterol level before and after doing exercise.

   (b) Compute the 99% confidence interval for the mean of the difference of cholesterol level before and after doing exercise.

   (c) According to the previous confidence intervals, what can you conclude about the effect of physical exercise on cholesterol?

2. In a pool performed in two hospital of a city patients were asked whether they were satisfied with the treatment. In the first hospital 200 patients were asked and 140 answered yes, while in the second hospital 300 patients were asked and 180 answered yes.

   (a) Compute the confidence interval for the difference of proportions of patients satisfied in each hospital.

   (b) Is there a significant difference, with a significance level 0.01, between the proportions of people satisfied in each hospital?

3. The data set neonates of the package rk.Teaching, contains information about a sample of 320 newborns that meet the normal gestation time in a hospital during one year. Do the following operations:

   (a) Compute the confidence interval for the difference between the weight means of neonates of non-smoker and smoker mothers. How much increase on average the weight of a newborn of a non-smoker mother compared to that of a smoker mother?

   (b) Considering only the sample of neonates of mothers that didn't smoke during the pregnancy, compute the confidence interval for the difference between the weight means of neonates of mothers that didn't smoke before the pregnancy and mothers that smoked before the pregnancy. Has the fact of smoking before the pregnancy a significant influence on the weight of neonates?

   (c) Compute the confidence interval for the mean of the difference between the APGAR scores at 1 minute and at 5 minutes. How is the evolution of neonates during the first minutes of life?

(d) If neonates with an APGAR score at 1 minute less than or equal to 3 are considered depressed, compute the 90% confidence interval for the difference between the proportions of depressed neonates with a smoker mother and with a non-smoker mother. Have the fact of smoking during the pregnancy a significant influence in the depression of neonates?

(e) Have the age of the mother a significant influence in the depression of neonates?