Statistics Practices with SPSS

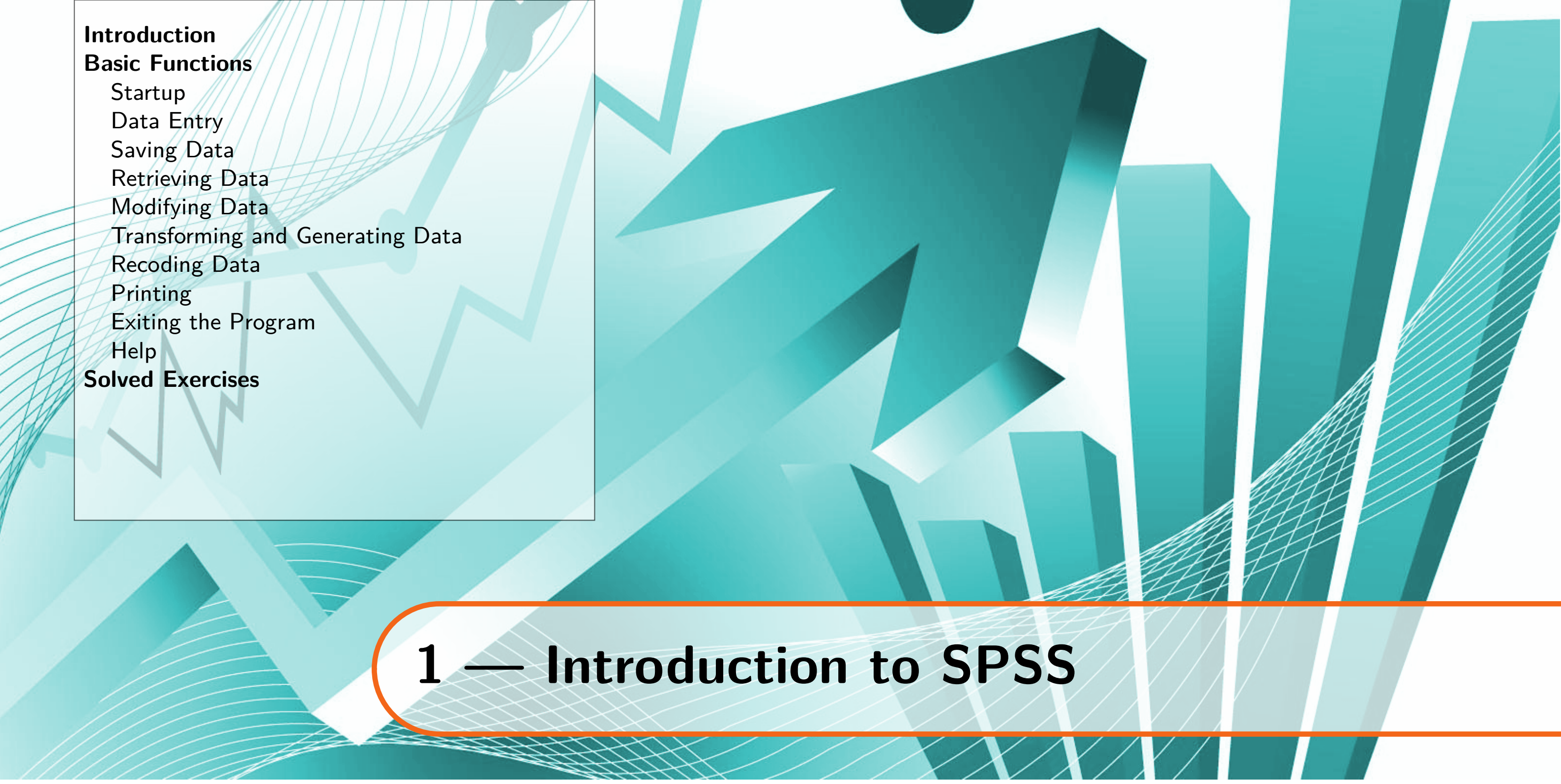
English Version 1
(February 9, 2025)

CEU
Universidad San Pablo

Prácticas de Estadística con SPSS
Santiago Angulo Díaz-Parreño, José Miguel Cárdenas Rebollo, Anselmo Romero Limón y Alfredo Sánchez Alberca.

Primera impresión, Abril 2013

Contents

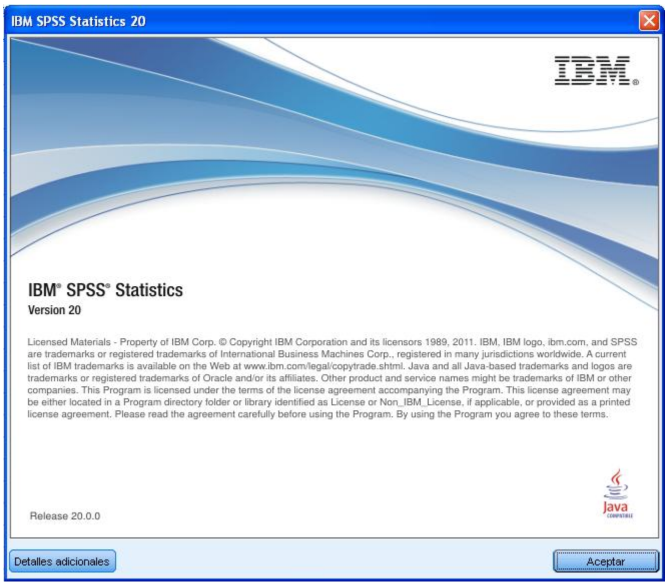


1 — Introduction to SPSS

1.1 Introduction

The great computational power achieved by computers has turned them into powerful tools at the service of all disciplines that, like statistics, require handling large volumes of data. Nowadays, practically no one considers conducting a serious statistical study without the help of a good statistical analysis program.

SPSS®* is one of the most widely used statistical analysis programs, especially in the field of biosciences.



IBM SPSS Statistics 20

IBM® SPSS® Statistics
Version 20

Licensed Materials - Property of IBM Corp. © Copyright IBM Corporation and its licensors 1989, 2011. IBM, IBM logo, ibm.com, and SPSS are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at www.ibm.com/legal/copytrade.shtml. Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates. Other product and service names might be trademarks of IBM or other companies. This Program is licensed under the terms of the license agreement accompanying the Program. This license agreement may be either located in a Program directory folder or library identified as License or Non_IBM_License, if applicable, or provided as a printed license agreement. Please read the agreement carefully before using the Program. By using the Program you agree to these terms.

Release 20.0.0

The objective of this practice is to introduce the student to the use of this program, teaching them to perform the most common basic operations. Throughout the practice, students will learn to create variables, enter sample data, transform variables, filter data, and merge and import data files.

*This practice is based on version 20.0 of SPSS® for Windows in Spanish/English.

1.2 Basic Functions

1.2.1 Startup

Like any other Windows application, to start the program, click on the corresponding option in the *Start*→*Programs* menu, or on the desktop icon

When the program starts, the data editor window appears (figure 1.1).

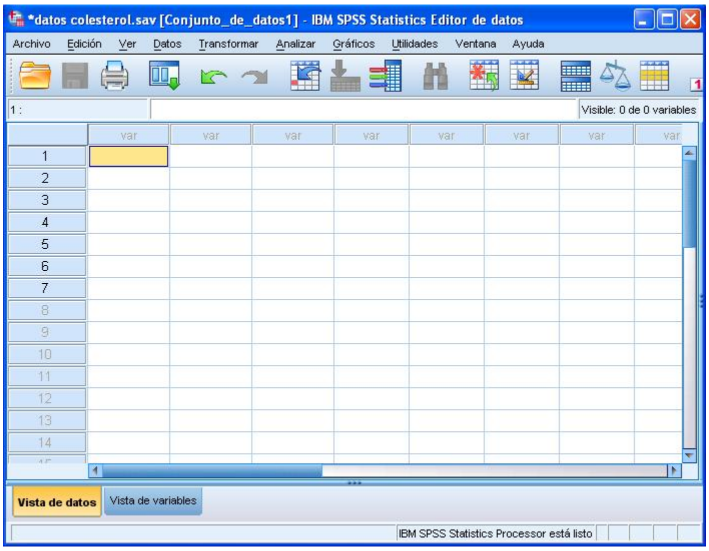

*datos colesterol.sav [Conjunto_de_datos1] - IBM SPSS Statistics Editor de datos
Archivo
Edición
Ver
Datos
Transformar
Analizar
Gráficos
Utilidades
Ventana
Ayuda
1 :
Visible: 0 de 0 variables
var
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Vista de datos
Vista de variables
IBM SPSS Statistics Processor está listo

Figure 1.1: Data editor window.

Like any other Windows application window, the data editor window has a title bar, a menu bar with the various functions SPSS can perform, including statistical data analysis, a toolbar with shortcuts to the most common menu options, and a status bar at the bottom that indicates what the program is doing at any given moment. Additionally, at the bottom, there are two tabs that allow switching to the *Data View* or the *Variable View*.

1.2.2 Data Entry

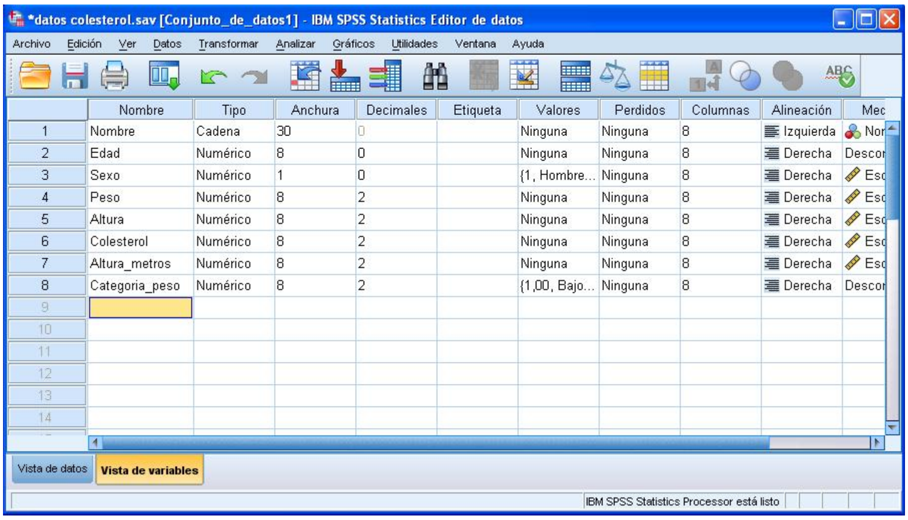
To perform any analysis, the data editor window must contain the data matrix to be analyzed. Once the user obtains the sample data, it must be entered into this window. To do this, the first step is to define the variables considered in the study. Each variable will correspond to a column in the data matrix.

To define a variable, we must switch to the *Variable View* by clicking on the corresponding tab (figure 1.2).

In this window, we must define each variable in a row, filling in the following fields:

Name The variable name can be any string of characters that begins with a letter and does not contain spaces or special characters such as ?,¿,*, etc. Each variable name must be unique, and case is not distinguished.

Type The most common types are Numeric (standard numeric format), Comma (with commas separating every three digits and a period for the decimal part), Dot (with dots separating every three digits and a comma for the decimal part), Scientific Notation (uses E for exponentiation), String (for alphanumeric data), and Date.


*datos colesterol.sav [Conjunto_de_datos1] - IBM SPSS Statistics Editor de datos
Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana Ayuda
Vista de datos
Vista de variables
IBM SPSS Statistics Processor está listo

	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Mec
1	Nombre	Cadena	30	0		Ninguna	Ninguna	8	Izquierda	Nor
2	Edad	Numérico	8	0		Ninguna	Ninguna	8	Derecha	Descor
3	Sexo	Numérico	1	0		{1, Hombre...	Ninguna	8	Derecha	Esc
4	Peso	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Esc
5	Altura	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Esc
6	Colesterol	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Esc
7	Altura_metros	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Esc
8	Categoria_peso	Numérico	8	2		{1,00, Bajo...	Ninguna	8	Derecha	Descor
9										
10										
11										
12										
13										
14										

Figure 1.2: Variable definition view.

Width This is the maximum number of characters that the variable values can have.

Decimals For numeric variables, this is the number of decimal places that can be written.

Label This is a description of the variable. If the variable name is sufficiently descriptive, this can be omitted.

Values This allows assigning labels to the different values that the variable can take. It is not mandatory but can be useful in some cases.

Clicking on the box opens a dialog box to assign value labels. To do this, simply enter a value in the *Value* text box and the corresponding label in the *Label* text box. Then click the *Add* button and repeat the same steps for all values of the variable. To finish, click the *OK* button.

Missing This allows defining which values will be used to represent missing data entered by the user. It is useful for distinguishing data that is missing for different reasons. For example, it may be interesting to distinguish missing data corresponding to a respondent who refuses to answer from missing data due to the question not being applicable to the respondent. User-specified missing values are excluded from most calculations.

Clicking on the box opens a dialog box where the discrete values representing missing values should be indicated (up to three can be entered), or the range of values that will be represented as missing.

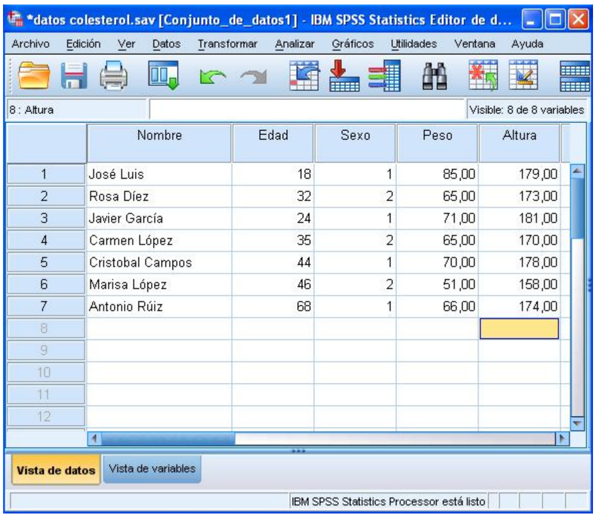
Columns This allows specifying the width of the column in which the data corresponding to the variable will be entered.

Alignment This allows specifying the alignment of the data corresponding to the variable. It can be Left, Right, or Centered.

Measure This allows specifying the type of scale used to measure the variable. It can be *Scale* when the variable is numeric and the scale is interval, *Ordinal* when the variable values represent categories with a certain order, or *Nominal* when the values represent unordered categories.

Role This allows specifying the role a variable has in the analysis. It can be *Input*, when it is an independent variable, *Target*, when it is a dependent variable, *Both*, when the variable can be both dependent and independent, *None* if the variable has no assigned role, *Partition*, when the variable will be used to divide the data into separate samples, and *Segment*, when it is a variable introduced to ensure compatibility in SPSS.

Once the variables are defined, the sample data is entered. To do this, return to the *Data View* by clicking on the corresponding tab. Now, the headers of the columns will show the names of the defined variables. Each individual in the sample corresponds to a row in the data matrix. To enter the value of a variable for a specific individual, we position ourselves in the cell of the row corresponding to that individual and the column of the variable, either by clicking on it or by navigating through the data matrix with the keyboard cursor arrows, and type the value followed by the *Enter* key (figure 1.3).


*datos colesterol.sav [Conjunto_de_datos1] - IBM SPSS Statistics Editor de d...
Archivo
Edición
Ver
Datos
Transformar
Analizar
Gráficos
Utilidades
Ventana
Ayuda

8 : Altura

Visible: 8 de 8 variables

	Nombre	Edad	Sexo	Peso	Altura
1	José Luis	18	1	85,00	179,00
2	Rosa Díez	32	2	65,00	173,00
3	Javier García	24	1	71,00	181,00
4	Carmen López	35	2	65,00	170,00
5	Cristobal Campos	44	1	70,00	178,00
6	Marisa López	46	2	51,00	158,00
7	Antonio Rúiz	68	1	66,00	174,00
8					
9					
10					
11					
12					

Vista de datos
Vista de variables
IBM SPSS Statistics Processor está listo

Figure 1.3: Entering data into the data matrix. Each column corresponds to a variable, and each row corresponds to an individual in the sample.

1.2.3 Saving Data

Once the data is entered, it is advisable to save it in a file to avoid having to re-enter it in future sessions. To do this, select the menu *File→Save*. If the file already exists, its information will be updated; if not, a dialog box will appear where you must enter the name you want to give the file and the folder where you want to save it. SPSS data files have the default extension *.sav. When the data is saved in a file, the file name will appear in the title of the data window (figure 1.3).

1.2.4 Retrieving Data

If the data you want to work with is already saved in a file, then you will need to open that file. To do this, select the menu *File→Open→Data* and select the file you want to open. Automatically, the data will appear in the data view.

1.2.5 Modifying Data

Sometimes it is necessary to modify the data in the data matrix to correct errors, add new data, or delete it. To correct a value, simply select the cell containing the value and type the new one. Other common operations are:

- Insert a new variable between existing ones. In the variable view, select the row containing the variable above which you want to insert the new one, and select the menu *Edit→Insert Variable*.
- Delete a variable. In the variable view, select the row containing the variable to be deleted and press the *Delete* key.
- Insert an individual between existing ones. In the data view, select the row containing the data of the individual above which you want to insert the new one, and select the menu *Edit→Insert Case*.

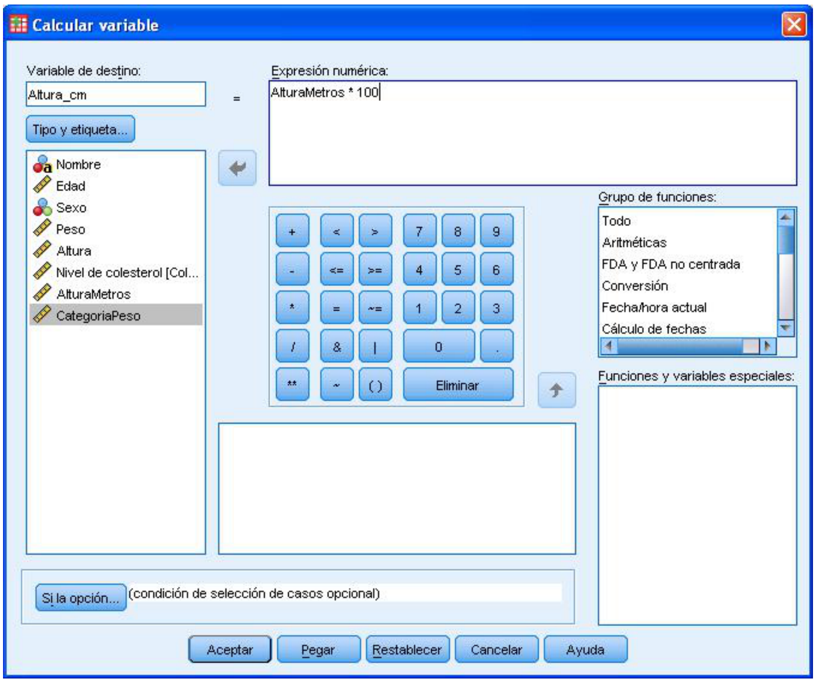
- Delete an individual. In the data view, select the row containing the data of the individual to be deleted and press the *Delete* key.

Whenever modifications are made to the data matrix, it is advisable to save the data again to update the file containing it.

Important!: When an undesired operation is performed by mistake, it can be undone using the menu *Edit*→*Undo*.

1.2.6 Transforming and Generating Data

In many statistical analyses, the data of the original variables is often transformed into more convenient forms for the analysis to be performed. To generate a new variable by transforming an existing one or using predefined functions, select the menu *Transform*→*Compute Variable...* Then the variable transformation window appears as shown in figure 1.4.



Calcular variable

Variable de destino: Altura_cm

Expresión numérica: AlturaMetros * 100

Figure 1.4: Variable transformation window. On the left are the already defined variables, on the right are the predefined functions that can be used, and in the center are the most common arithmetic and relational operators.

In this window, you must enter the name of the new variable in the *Target Variable* box, and the expression whose result will be the content of the new variable in the *Numeric Expression* box. For this, a series of operators and functions are available to perform the transformation, as well as the list of already defined variables that can be used as arguments for the various transformation functions.

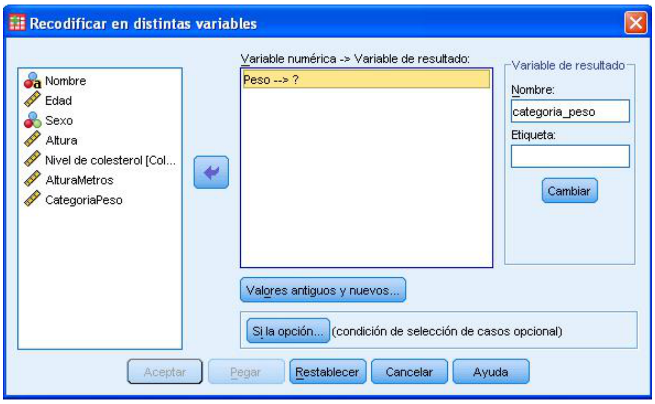
The most common operators for building expressions are the arithmetic ones +, -, *, /, ** (exponentiation), the relational ones =, <, >, ~=, <=, >=, and the logical ones & (AND), | (OR), and ~ (negation). Some of the most common functions are: ABS (absolute value), SQRT (square root), EXP (exponential), LN (natural logarithm), SIN (sine), COS (cosine), TAN (tangent), SUM (sum), MEAN (arithmetic mean), SD (standard deviation), RND (rounding to the nearest integer), TRUNC (integer part of a number).

Clicking on the *If...* button allows setting conditions for applying the transformation. To set a condition, you must activate the option *Include if case satisfies condition* and then enter a logical condition such as Sex=1. In this way, the transformation will only be applied to individuals who meet this condition.

Once the expression is defined, click on the *OK* button, and automatically a new column with the transformed data of the new variable will appear in the data view.

1.2.7 Recoding Data

Another way to transform a variable is to create another whose values are a recoding of the first, for example by grouping into intervals. This recoding can be done either in the same variable or in different variables. To do this, select the menu *Transform→Recode into Different Variables*. Automatically, the variable recoding window appears as seen in figure 1.5.



Recodificar en distintas variables

Nombre
Edad
Sexo
Altura
Nivel de colesterol [Col...
AlturaMetros
CategoriaPeso

Variable numérica -> Variable de resultado: Peso --> ?

Variable de resultado

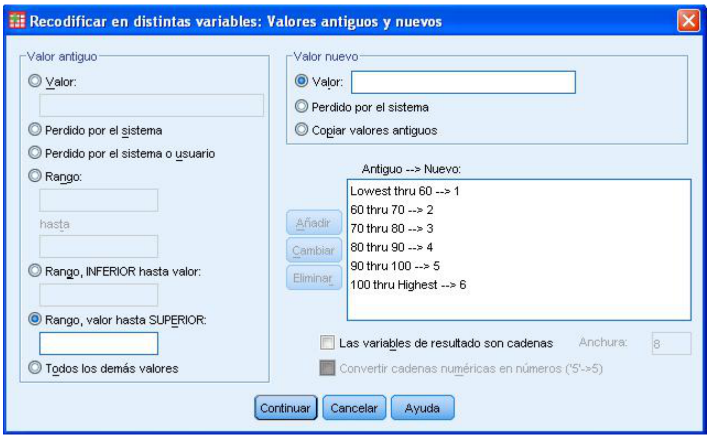
Nombre: categoria_peso

Etiqueta:

Cambiar
Valores antiguos y nuevos...
Si la opción...
(condición de selección de casos opcional)
Aceptar
Pegar
Restablecer
Cancelar
Ayuda

Figure 1.5: Variable recoding window. On the left are the already defined variables, on the right the recoding rules must be specified.

To recode a variable into a new one, first select the variable you want to recode and click on the button with an arrow next to it. Then enter the name of the new variable in the *Name* box and click on the *Change* button. Next, set the recoding rules. To do this, click on the *Old and New Values* button to bring up the rule definition window (figure 1.6). The rules can establish the conversion of the value of the original variable entered in the *Old Value* box into the value of the new variable entered in the *New Value* box, or the conversion of an entire range of values of the original variable into a value of the new variable. Once these values are defined, click on the *Continue* button, and then on *OK*.



Recodificar en distintas variables: Valores antiguos y nuevos

Valor antiguo

○ Valor:

○ Perdido por el sistema

○ Perdido por el sistema o usuario

○ Rango:

hasta

○ Rango, INFERIOR hasta valor:

● Rango, valor hasta SUPERIOR:

○ Todos los demás valores

Valor nuevo

● Valor:

○ Perdido por el sistema

○ Copiar valores antiguos

Antiguo --> Nuevo:

Lowest thru 60 --> 1
60 thru 70 --> 2
70 thru 80 --> 3
80 thru 90 --> 4
90 thru 100 --> 5
100 thru Highest --> 6

□ Las variables de resultado son cadenas Anchura: 8

□ Convertir cadenas numéricas en números ('5'->5)

Figure 1.6: Rule definition window.

1.2.8 Printing

To print, use the menu *File→Print*. A print dialog box immediately appears where you must indicate whether you want to print everything or just the selection you have made. After this, click on the *OK* button, and the information is sent to the printer.

Before printing, it is advisable to preview what will be sent to the printer to ensure that it is what you want. To do this, use the menu *File→Print Preview*. Then a viewer appears where you can see the page as it will be sent to the printer. If everything looks correct, you can click on the *Print* button, and the print dialog box will appear from which you can send it to the printer definitively.

1.2.9 Exiting the Program

To end a work session, use the menu *File→Exit*, or click on the cross to close the program window. If there is data or results that have not been saved, the program will ask you before exiting if you want to save them.

1.2.10 Help

In this practice, only the basic operations in a work session have been described. But all the statistical analyses that can be performed with the menus in the menu bar remain to be described. Although many of these menus will be explained in the following practices, the program has the help menu *Help* where you can find a description of all these menus and to which you can turn whenever you have doubts.

1.3 Solved Exercises

1. Enter the data of the following sample into the data matrix and save it in a file named cholesterol_data.sav.

Name	Sex	Weight	Height	Cholesterol
José Luis Martínez Izquierdo	M	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
Antonio Fernández Ocaña	M	62	172	276
Pilar Martín González	F	60	166	213
Pedro Gálvez Tenorio	M	90	194	241
Santiago Reillo Manzano	M	75	185	280
Macarena Álvarez Luna	F	55	162	262
José María de la Guía Sanz	M	78	187	198
Miguel Angel Cuadrado Gutiérrez	M	109	198	210
Carolina Rubio Moreno	F	61	177	194

(a) In the *Variable View* window, create the variables Name, Sex, Weight, Height, and Cholesterol and enter the above data, following the instructions in section 1.2.2.

(b) Once the data is entered, save it in a file named cholesterol_data following the instructions in section 1.2.3.

2. Perform the following operations on the data matrix from the previous exercise:

(a) Insert a new variable Name with the ages of all individuals in the sample.

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
Antonio Fernández Ocaña	51
Pilar Martín González	22
Pedro Gálvez Tenorio	35
Santiago Reillo Manzano	46
Macarena Álvarez Luna	53
José María de la Guía Sanz	58
Miguel Angel Cuadrado Gutiérrez	27
Carolina Rubio Moreno	20

i. In the *Variable View*, select the row corresponding to the variable Sex by clicking on its header, then select the menu *Edit→Insert Variable*, which will insert a new row between the variables Name and Sex.

ii. In the new row, define the variable Age and enter the previous data.

iii. In the *Data View*, fill in the data for the column corresponding to the variable Age.

(b) Insert the following individual's data between the 4th and 5th individuals:

Name: Cristóbal Campos Ruiz.
Age: 44 years.

Sex: Male.
Weight: 70 Kg.
Height: 178 cm.
Cholesterol: 220 mg/dl.

> i. Select the row corresponding to the 5th individual by clicking on its header, then select the menu *Edit→Insert Case*, which will insert a new row between the 4th and 5th individuals.
> ii. Enter the indicated data in the new row.

(c) Change Macarena Álvarez Luna's weight value to 58.

> Click on the cell you want to modify, type 58, and press *Enter*.

(d) Transform the variable Height to be expressed in meters.

> i. Select the menu *Transform→Compute Variable*.
> ii. In the data transformation window, enter the name Height meters in the box for *Target Variable*.
> iii. Enter the expression Height/100 in the box for *Numeric Expression*.
> iv. Click on the button *OK*.

(e) Recode the variable weight into four categories, considering sex:

Category	Men	Women
Low	≤ 70	≤ 60
Medium	(70,85]	(60,70]
High	(85,100]	(70,80]
Very High	> 100	> 80

> i. Select the menu *Transform→Recode into Different Variables*.
> ii. In the recoding window, select the variable Weight and click on the arrow button next to it.
> iii. Enter the name of the recoded variable as Weight Category in the box for *Name* under *Output Variable*, and click on the button *Change*.
> iv. Click on the button *Old and New Values* to open the recoding rules definition window.
> v. To define recoding rules for men,
>   A. Select the option *Range LOWEST through value* under *Old Value*, enter 70 in the box. Enter 1 under *New Value*, and click on *Add*.
>   B. Select the option *Range* under *Old Value*, enter 70 and 85 in the boxes. Enter 2 under *New Value*, and click on *Add*.
>   C. Select the option *Range* under *Old Value*, enter 85 and 100 in the boxes. Enter 3 under *New Value*, and click on *Add*.
>   D. Select the option *Range value through HIGHEST* under *Old Value*, enter 100 in the box. Enter 4 under *New Value*, and click on *Add*.
> vi. Click on the button *Continue* to close the window.
> vii. Click on the button *If...* to open the condition definition window.
> viii. Select the option *Include if case satisfies condition*, enter the condition Sex="M" in the box.
> ix. Click on the button *Continue* to close the window.

x. Click on the button *OK*.

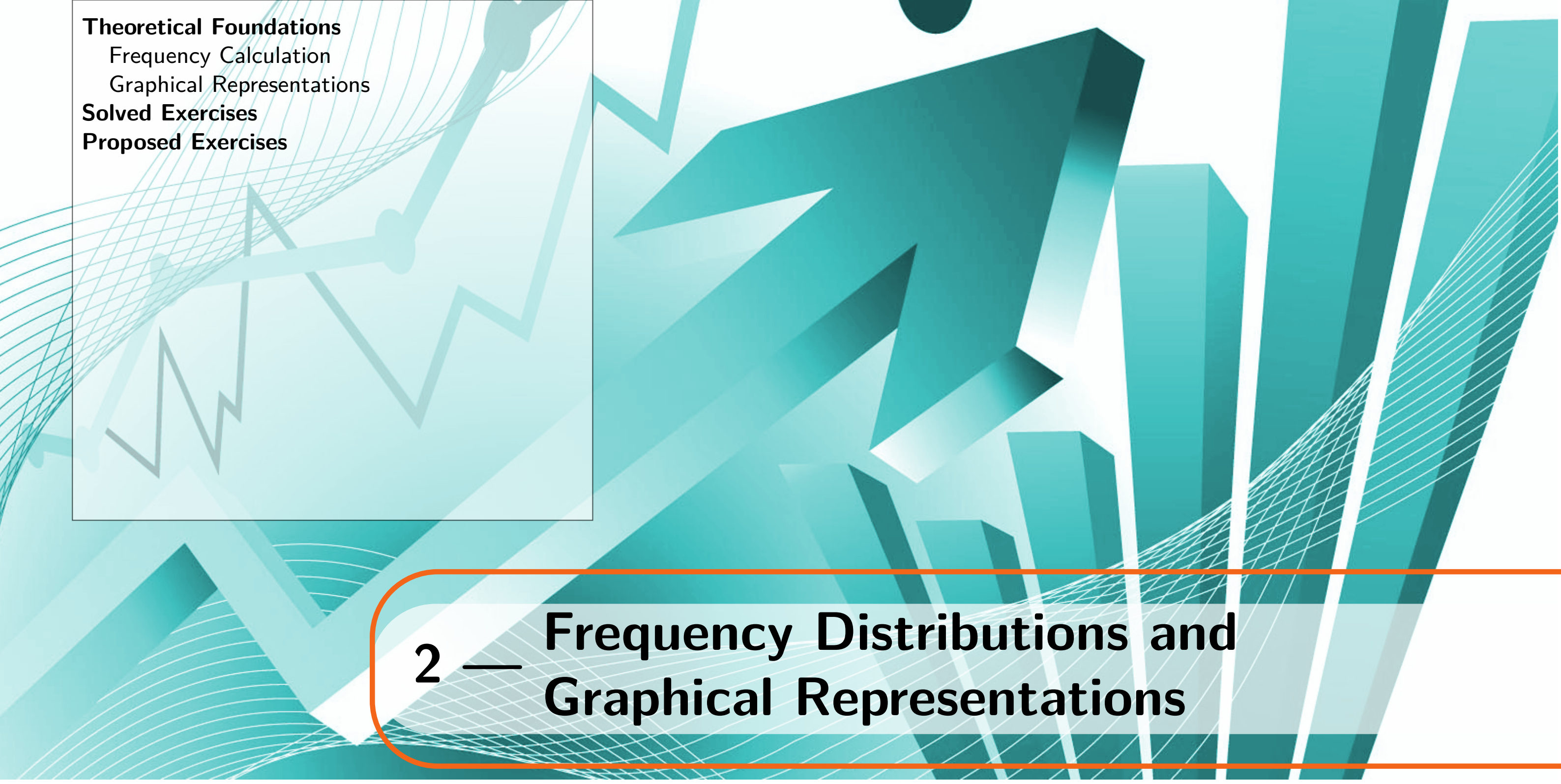
xi. Repeat the same steps to define recoding rules for women.

xii. In *Variable View*, click on *Values of Weight Category*, and in *Value Labels* assign labels Low, Medium,High, and Very High to values 1, 2, 3, and 4 respectively, clicking on *Add* after each assignment, and finally click on *OK*.

(f) Save changes to the previous file and exit the program.

i. Select the menu *File*→*Save*.

ii. Select the menu *File*→*Exit*.



2 — Frequency Distributions and Graphical Representations

2.1 Theoretical Foundations

One of the first steps in any statistical study is the summary and description of the information contained in a sample. For this, some descriptive analysis methods will be applied, which will allow us to classify and structure the information as well as represent it graphically.

The characteristics we study may or may not be measurable; in this sense, we will define a *variable* as a characteristic that can be measured, that is, quantitative and quantifiable through observation (for example, people's weight, age, etc.), and we will define an *attribute* as a characteristic that cannot be measured, and consequently observable only qualitatively (for example, eye color, a patient's condition, etc.). The possible observations of an attribute are called modalities.

Within the attributes, we can talk about *ordinal attributes*, those that present some type of order among the different modalities, and *nominal attributes*, in which there is no order among them.

Within the variables, we can differentiate between *discrete*, if their possible values are isolated values, and *continuous*, if they can take any value within an interval.

In some texts, the term *attribute* is not used, and all characteristics are called *variables*. In that case, *quantitative variables* are distinguished to designate those we have defined here as *variables*, and *qualitative variables* for those we have called *attributes*. Henceforth, this criterion will be applied to simplify the exposition.

2.1.1 Frequency Calculation

To study any characteristic, the first thing we must do is count the observations and the number of repetitions of these. For each value x_i of the sample, it is defined:

Absolute frequency It is the number of times each of the values x_i appears and is denoted by n_i.

Relative frequency It is the number of times each value x_i appears divided by the sample size and is denoted by f_i

$$f_i = \frac{n_i}{n}$$

Generally, relative frequencies are multiplied by 100 to represent the percentage.

In the case that there is an order among the values of the variable, sometimes we are interested not only in knowing the number of times a certain value is repeated but also the number of times that value and all previous ones appear. This type of frequency is called *cumulative frequencies.*

Cumulative absolute frequency It is the sum of the absolute frequencies of the values less than x_i plus the absolute frequency of x_i, and is denoted by N_i

$$N_i = n_1 + n_2 + \ldots + n_i$$

Cumulative relative frequency It is the sum of the relative frequencies of the values less than x_i plus the relative frequency of x_i, and is denoted by F_i

$$F_i = f_1 + f_2 + \ldots + f_i$$

The results of the observations of the values of a statistical variable in a sample are usually represented in the form of a table. In the first column, the values x_i of the variable are represented in ascending order, and in the next column, the values of the corresponding absolute frequencies n_i.

We can complete the table with other columns, corresponding to the relative frequencies, f_i, and the cumulative frequencies, N_i and F_i. The set of values of the variable observed in the sample along with their frequencies is known as the *sample frequency distribution.*

■ **Example 2.1** In a survey of 25 couples about the number of children they have, the following data are obtained: 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. The distinct values of the variable are: 0, 1, 2, 3, and 4. Thus, the absolute frequency would be:

x_i	*Count*	n_i
0	*II*	2
1	*IIIIII*	6
2	*IIIIIIIIIIIIII*	14
3	*II*	2
4	*I*	1

And the frequency distribution table would be:

x_i	n_i	f_i	N_i	F_i
0	2	0.08	2	0.08
1	6	0.24	8	0.32
2	14	0.56	22	0.88
3	2	0.08	24	0.96
4	1	0.04	25	1
Sum	25	1		

■

When the sample size is large in the case of discrete variables with many distinct values of the variable, and in any case if it is continuous variables, the observations are grouped into *classes*, which are contiguous intervals, preferably of the same width.

To decide the number of classes to consider, a frequently used rule is to take the integer closest to $\sqrt{n}$ where n is the number of observations in the sample. But it is advisable to try different numbers of classes and choose the one that provides a clearer description. Thus, the intervals $(a_{i-1}, a_i], i = 1, 2, \ldots, l$ are prefixed, being $a = a_0 < a_1 < .... < a_l = b$ so that all observed values are within the interval $(a, b]$, and without ambiguity in deciding which interval each data belongs to.

We will call *class mark* the midpoint of each interval. Thus, the *class mark* $(a_{i-1}, a_i]$ is the midpoint x_i of that class, that is

$$x_i = \frac{a_{i-1} + a_i}{2}$$

In the statistical treatment of grouped data, all values in the same class are considered equal to the class mark. In this way, if in the class $(a_{i-1}, a_i]$ there are n_i observed values, the class mark x_i can be associated with this frequency n_i.

2.1.2 Graphical Representations

We have seen that the statistical table summarizes the data of a sample, so that it can be analyzed in a more systematic and summarized way. To achieve a visual perception of the characteristics of the population, the use of graphs and diagrams is very useful. Depending on the type of variable and whether we work with grouped data or not, different types will be used.

Bar Chart and Frequency Polygon

It consists of representing on the x-axis of a coordinate system the different values of the variable X, and raising a bar on each of those points whose height is equal to the corresponding absolute or relative frequency of that value, as shown in figure 2.1(a). This representation is used for frequency distributions with few distinct values of the variable, both quantitative and qualitative, and in the latter case, it is usually represented with rectangles of height equal to the frequency of each modality.

In the case of quantitative variables, the bar chart of cumulative frequencies can also be represented, as shown in figure 2.1(b).

Another common representation is the *frequency polygon* which consists of the polygonal line whose vertices are the points (x_i, n_i), as seen in figure 2.1(c), and if instead of considering the absolute or relative frequencies, the cumulative absolute or relative frequencies are considered, the *cumulative frequency polygon* is obtained, as seen in figure 2.1(d).

Histograms

This type of representation is used for continuous variables and discrete variables where observations have been grouped into classes. A *histogram* is a set of rectangles, whose bases are the class intervals $(a_{i-1}, a_i]$ on the OX axis and their height the corresponding absolute, relative, cumulative absolute, or cumulative relative frequency, as shown in figures 2.2(a) and 2.2(b).

If we join the midpoints of the upper bases of the rectangles of the histogram, we obtain the *frequency polygon* corresponding to grouped data (figure 2.2(c)).

The frequency polygon can also be used to represent cumulative frequencies, both absolute and relative. In this case, the polygonal line is drawn by joining the right ends of the upper bases of the rectangles of the cumulative frequency histogram, instead of the central points (figure 2.2(d)).

For qualitative and discrete quantitative variables, representative surfaces can also be used; among these, the most commonly used are *pie charts*.

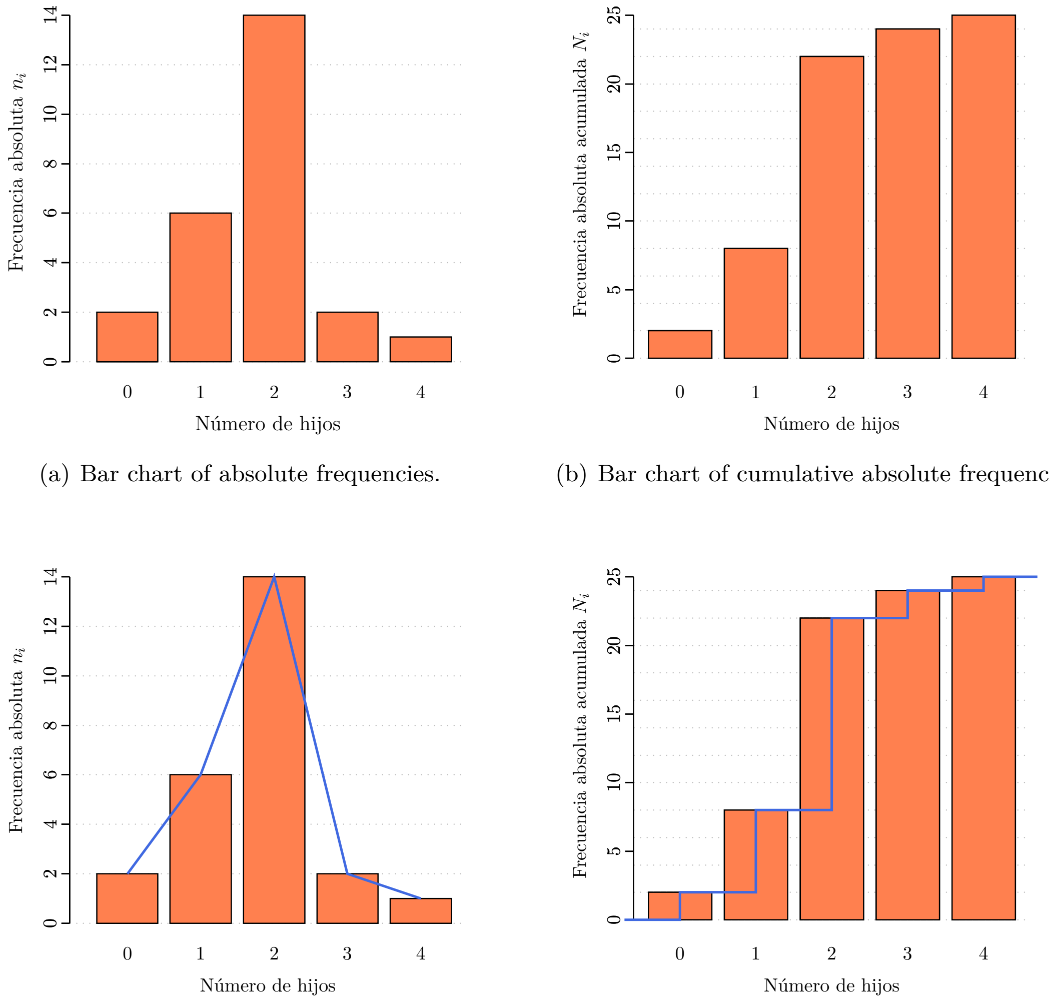
Pie Charts

It is a representation in which a circle is divided into sectors, so that the angles, and therefore the respective areas, are proportional to the frequency.

■ **Example 2.2** A study is being conducted in a population on the blood group of its citizens. For this, we have a sample of 30 people, with the following results: 5 people with group 0, 14 with group A, 8 with group B, and 3 with group AB. The pie chart of relative frequencies corresponding to this appears in figure 2.3. ■

Box Plot and Outliers

Extremely high or low data, compared to the rest of the sample, are called influential data or *outliers*. Such data, as their name suggests, can modify the conclusions of a study and should be carefully considered before accepting them, as they may often simply be erroneous



(a) Bar chart of absolute frequencies.

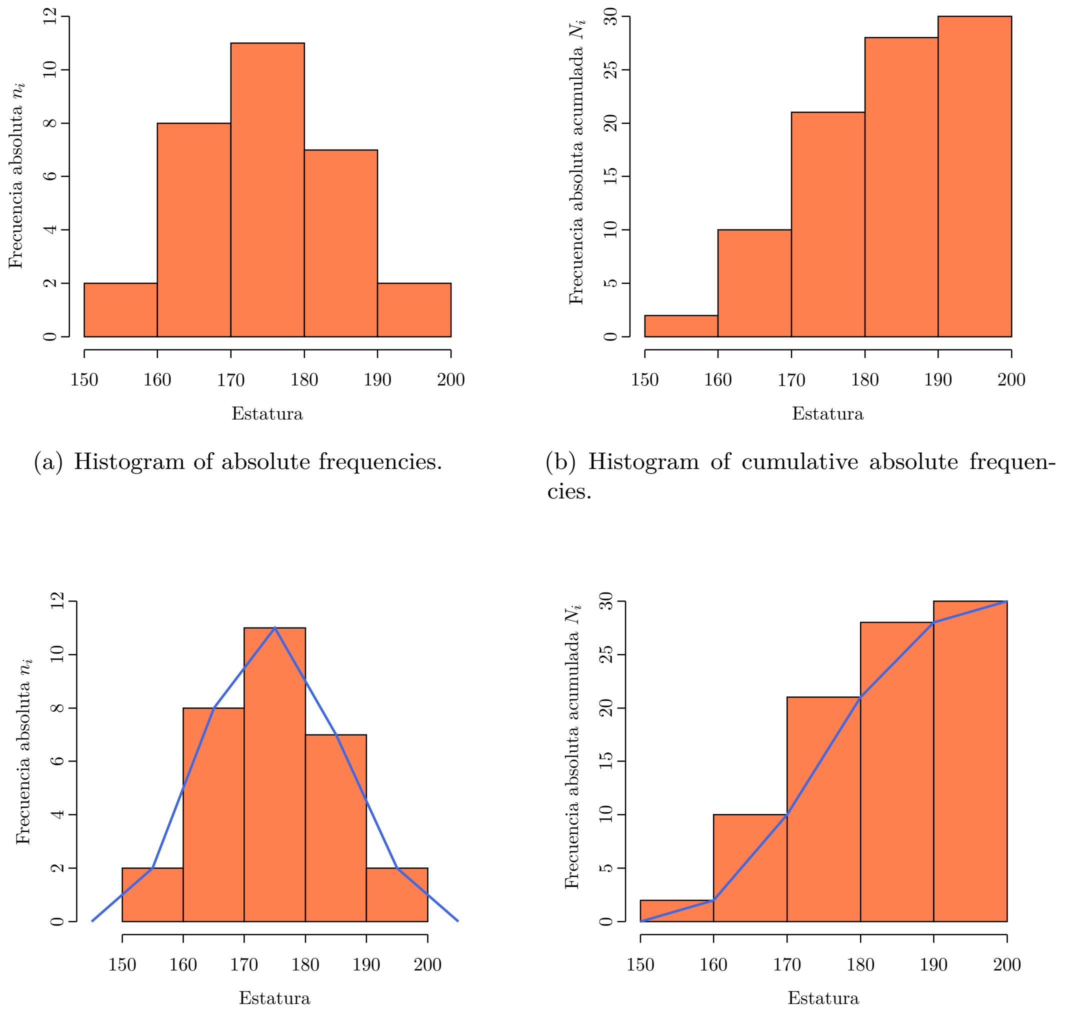
(b) Bar chart of cumulative absolute frequencies.

(c) Frequency polygon of absolute frequencies.

(d) Cumulative frequency polygon

Figure 2.1: Bar charts and associated polygons for ungrouped data.

data. The most appropriate graphical representation to detect these data is the *box plot*. This diagram consists of a box that contains 50% of the central data of the distribution, and segments that come out of the box, indicating the limits beyond which the data are considered outliers. In figure 2.4, an example can be seen in which two outliers appear.



(a) Histogram of absolute frequencies.

(b) Histogram of cumulative absolute frequencies.

(c) Frequency polygon of absolute frequencies.

(d) Cumulative frequency polygon

Figure 2.2: Histogram and associated polygons for grouped data.

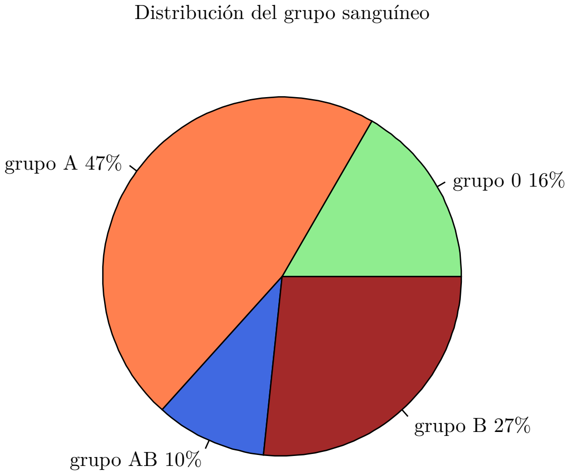


Figure 2.3: Pie chart of relative frequencies of blood group.

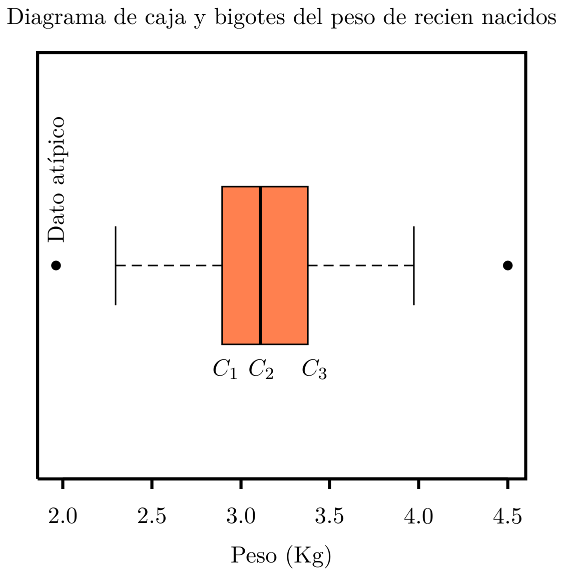


Figure 2.4: Box plot for a sample of newborns. There are two children with outlier weights, one with an extremely low weight of 1.9 kg, and another with an extremely high weight of 4.5 kg.

2.2 Solved Exercises

(a) A survey was conducted with 40 people over 70 years old about the number of different medications they usually take. The result of this survey was as follows:

$$3-1-2-2-0-1-4-2-3-5-1-3-2-3-1-4-2-4-3-2$$
$$3-5-0-1-2-0-2-3-0-1-1-5-3-4-2-3-0-1-2-3$$

i. Create the variable medications and enter the data.
ii. Construct the frequency table.

> A. Select the menu *Analyze→Descriptive Statistics→Frequencies*.
> B. Select the variable medications in the *Variables* field of the dialog box.
> C. Activate the option *Display frequency tables* and click the *OK* button.

iii. Draw the bar chart of absolute frequencies.

> A. Select the menu *Graphs→Legacy Dialogs→Bar→Define*.
> B. Select the variable medications in the *Category Axis* field of the dialog box and select the option *N of cases*.

iv. Draw the frequency polygon of absolute frequencies.

> A. Select the menu *Graphs→Legacy Dialogs→Line→Define*.
> B. Select the variable medications in the *Category Axis* field of the dialog box and select the option *N of cases*.

v. Draw the bar chart of cumulative relative frequencies.

> Repeat the same steps as in section c) but this time select the option *% of cases*.

vi. Draw the pie chart.

> A. Select the menu *Graphs→Legacy Dialogs→Pie→Define*.
> B. Select the variable medications in the *Define Slices by* field of the dialog box.

(b) A study was conducted in a hospital on the number of people admitted to the emergency room in the month of November. The observed data were:

$$15-23-12-10-28-7-12-17-20-21-18-13-11-12-26$$
$$29-6-16-39-22-14-17-21-28-9-16-13-11-16-20$$

i. Create the variable emergencies and enter the data.
ii. Draw the histogram of absolute frequencies grouped into 5 classes from 0 to 40.

> A. Select the menu *Graphs→Legacy Dialogs→Histogram*.
> B. Select the variable emergencies in the *Variable* field of the dialog box and click the *OK* button.
> C. Edit the histogram by double-clicking on it.
> D. In the chart editor, right-click in the histogram area, which will be surrounded by a yellow line, and click on *Properties Window* in the pop-up window.
> E. Select the option *Bin Sizes*, in *X Axis* choose *Custom* and in *Number of*

Intervals put 5.

F. Click the *Apply* button, then the *Close* button in the properties window, and close the chart editor.

iii. Draw the box plot. Are there any outliers?

> ℹ
>
> A. Select the menu *Graphs→Legacy Dialogs→Boxplot*.
>
> B. Select the option *Summaries for groups of cases* and click the *Define* button.
>
> C. Select the variable emergencies in the *Boxes represent* field of the dialog box and click the *OK* button.

iv. If there are any outliers, remove them and draw the histogram of absolute frequencies, so that classes of width 5 appear, starting at 5 and ending at 30.

> ℹ
>
> A. Identify the case corresponding to the outlier and remove it in the data editor.
>
> B. Repeat the steps in section b) to draw the histogram.

2.3 Proposed Exercises

(a) The number of injuries suffered during a season by each player of a soccer team was as follows:

$$0-1-2-1-3-0-1-0-1-2-0-1$$
$$1-1-2-0-1-3-2-1-2-1-0-1$$

Tasks:

i. Create the variable injuries and enter the data.

ii. Construct the frequency table.

iii. Draw the bar chart of cumulative relative frequencies.

iv. Draw the frequency polygon of cumulative absolute frequencies.

v. Draw the pie chart.

(b) To conduct a study on the height of university students, we selected, through a random sampling process, a sample of 30 students, obtaining the following results (measured in centimeters):

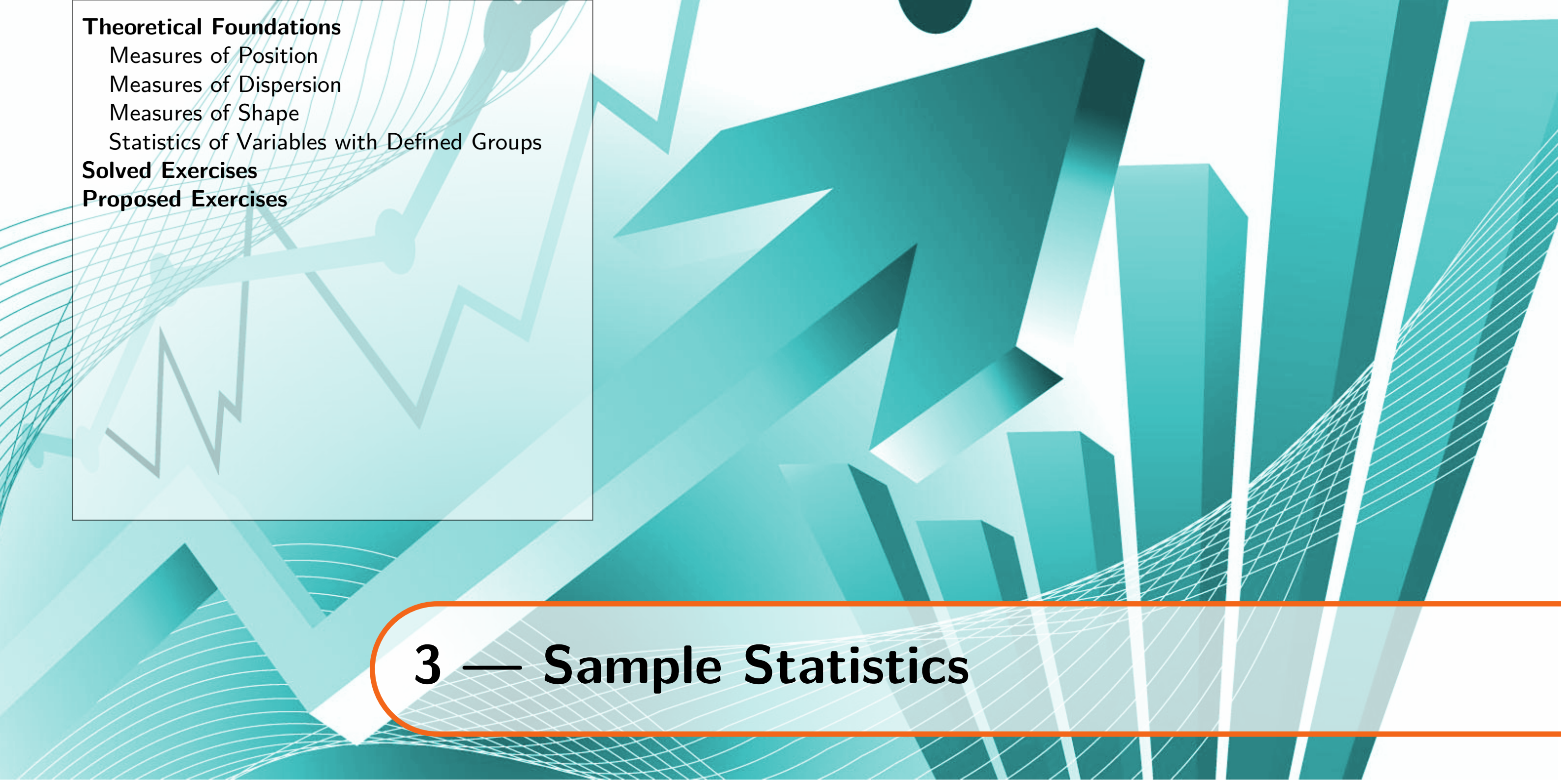
179, 173, 181, 170, 158, 174, 172, 166, 194, 185,
162, 187, 198, 177, 178, 165, 154, 188, 166, 171,
175, 182, 167, 169, 172, 186, 172, 176, 168, 187.

Tasks:

i. Create the variable height and enter the data.

ii. Draw the histogram of absolute frequencies grouped from 150 to 200 in classes of width 10.

iii. Draw the box plot. Are there any outliers?



3 — Sample Statistics

3.1 Theoretical Foundations

We have seen how we can present the information obtained from the sample through tables or graphs. The frequency table contains all the information of the sample, but it is difficult to draw conclusions about certain aspects of the distribution just by looking at it. Now we will see how, from these same observed values of the statistical variable, certain numbers are calculated that summarize the sample information. These numbers, called *Statistics*, are used to highlight certain aspects of the distribution, such as the dispersion or concentration of the data, the shape of their distribution, etc. Depending on the characteristic they aim to reflect, they can be classified into Measures of Position, Measures of Dispersion, and Measures of Shape.

3.1.1 Measures of Position

These are values that indicate how the data are positioned. The most important ones are the Arithmetic Mean, the Median, and the Mode.

Arithmetic Mean $\overline{x}$

The *arithmetic mean* of a statistical variable X is denoted by $\overline{x}$ and is defined as the sum of all observed results divided by the sample size. That is, the mean of the statistical variable X, whose frequency distribution is (x_i, n_i), is given by

$$\overline{x} = \frac{x_1 + \ldots + x_1 + \ldots + x_k + \ldots + x_k}{n_1 + \ldots + n_k} = \frac{x_1 n_1 + \ldots + x_k n_k}{n} = \frac{1}{n} \sum_{i=1}^{k} x_i n_i$$

The arithmetic mean only makes sense for quantitative variables.

Median Me

The *median*, denoted by Me, is the value of the sample that, once all the values are arranged in ascending order, has as many terms below it as above it. Consequently, it divides the distribution into two equal parts.

The median only makes sense for ordinal attributes and quantitative variables.

Mode *Mo*

The *mode* is the value of the variable that appears most frequently in the sample. When there is more than one value with the maximum frequency, we say there is more than one mode. For continuous or grouped discrete variables, we call the class with the maximum frequency the modal class. The mode can be calculated for both quantitative and qualitative variables.

Quantiles

If the total set of observed values is divided into r parts, each containing $\frac{n}{r}$ observations, the separation points are generically called *quantiles.*

According to this, the median is also a quantile with $r = 2$. Some quantiles have specific names, such as:

Quartiles. These are the points that divide the distribution into 4 parts, with an equal number of observations in each, and are denoted by C_1, C_2, C_3. It is clear that $C_2 = Me$.

Deciles. These are the points that divide the distribution into 10 parts, with an equal number of observations in each, and are denoted by $D_1, D_2, \ldots, D_9$.

Percentiles. These are the points that divide the distribution into 100 parts, with an equal number of observations in each, and are denoted by $P_1, P_2, \ldots, P_{99}$.

3.1.2 Measures of Dispersion

These measure the spread between the values in the sample. The most important ones are the Range, Interquartile Range, Variance, Standard Deviation, and Coefficient of Variation.

Range *Re*

The most immediate measure of dispersion is the range. We call the *range* and denote it by *Re* the difference between the maximum and minimum values of the variable in the sample. That is

$$Re = max\{x_i, i = 1, 2, \ldots, n\} - min\{x_i, i = 1, 2, \ldots, n\}$$

This statistic measures the range of variation of the variable, although it provides the least information about the clustering of the values around the measures of central tendency. Additionally, it has the disadvantage of being highly affected by outliers.

Interquartile Range *RI*

The *interquartile range RI* is the difference between the third and first quartiles, and thus measures the range of variation of the central 50% of the data in the distribution. Therefore

$$RI = C_3 - C_1$$

The advantage of the interquartile range over the range is that it is less affected by outliers.

Variance s_x^2

We call the *variance* of a statistical variable X, denoted by s_x^2, the mean of the squares of the deviations of the observed values from the sample mean. Thus

$$s_x^2 = \frac{1}{n} \sum_{i=1}^{k} (x_i - \overline{x})^2 n_i$$

Standard Deviation s_x

The positive square root of the variance is known as the *standard deviation* of the variable X, and is denoted by s

$$s = +\sqrt{s_x^2}$$

Pearson's Coefficient of Variation *Cv_x*

The ratio between the standard deviation and the absolute value of the mean is known as *Pearson's coefficient of variation* or simply *coefficient of variation*:

$$Cv_x = \frac{s_x}{|\overline{x}|}$$

The coefficient of variation is dimensionless, and thus allows for comparisons between variables expressed in different units. The closer it is to 0, the lower the dispersion of the sample relative to the mean, and the more representative the mean is of the set of observations.

3.1.3 Measures of Shape

These indicate the shape of the distribution of values in the sample. They can be classified into two groups: Measures of *skewness* and measures of *kurtosis*.

Fisher's Skewness Coefficient *g₁*

The *Fisher's skewness coefficient*, denoted by g_1, is defined as

$$g_1 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^3 f_i}{s_x^3}$$

Depending on its value, we have:

- $g_1 = 0$. Symmetrical distribution.
- $g_1 < 0$. Left-skewed distribution.
- $g_1 > 0$. Right-skewed distribution.

Kurtosis Coefficient *g₂*

The degree of peakedness of the sample observations is characterized by the *kurtosis coefficient* and is denoted by g_2

$$g_2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^4 f_i}{s_x^4} - 3$$

Depending on its value, we have:

- $g_2 = 0$. The distribution has the same peakedness as a normal distribution with the same mean and standard deviation. It is said to be a *mesokurtic* distribution.
- $g_2 < 0$. The distribution is less peaked than a normal distribution with the same mean and standard deviation. It is said to be a *platykurtic* distribution.
- $g_2 > 0$. The distribution is more peaked than a normal distribution with the same mean and standard deviation. It is said to be a *leptokurtic* distribution.

Both g_1 and g_2 are often used to check if the sample data come from a non-normal population. When g_1 is outside the interval [-2,2], the distribution is considered too skewed for the data to come from a normal population. Similarly, when g_2 is outside the interval [-2,2], the distribution is considered either too peaked or too flat for the data to come from a normal population.

3.1.4 Statistics of Variables with Defined Groups

We already know how to summarize the information contained in a sample using a series of statistics. But so far, we have only studied examples with a single character under study.

In most research, we will not study a single character, but a set of characters, and often it will be convenient to obtain information about a specific character based on the groups created by another character studied in the research. These variables used to form groups are known as *classifying* or *discriminant* variables.

For example, if a study is conducted on a group of newborn children, we can study their weight. But if we also know whether each child's mother is a smoker or not, we can study the weight of children of smoking mothers on one hand and non-smoking mothers on the other, to see if there are differences between the two groups.

3.2 Solved Exercises

(a) A survey was conducted on 40 people over 70 years old about the number of different medications they regularly take. The results of the survey were as follows:

$$3-1-2-2-0-1-4-2-3-5-1-3-2-3-1-4-2-4-3-2$$
$$3-5-0-1-2-0-2-3-0-1-1-5-3-4-2-3-0-1-2-3$$

The following is requested:

i. Create the variable medications and enter the data. If the data is already available, simply retrieve it.

ii. Calculate the arithmetic mean, median, mode, variance, and standard deviation of this variable. Interpret the statistics.

> A. Select the menu *Analyze→Descriptive Statistics→Frequencies*.
> B. Select the variable medications in the *Variables* field of the dialog box.
> C. Click on the *Statistics* button. To select only the requested statistics, check the boxes corresponding to these statistics and click on the *Continue* and *OK* buttons.

iii. Calculate the skewness and kurtosis coefficients and interpret the results.

> Follow the same steps as in the previous section, now selecting the requested statistics.

iv. Calculate the quartiles.

> Follow the same steps as in the previous sections, activating the *Quartiles* option.

(b) In a group of 20 students, the grades obtained in Mathematics were:

SS - AP - SS - AP - AP - NT - NT - AP - SB - SS
SB - SS - AP - AP - NT - AP - SS - NT - SS - NT

i. Create the variable grades and enter the data.

ii. Recode this variable, assigning 2.5 to SS, 5.5 to AP, 7.5 to NT, and 9.5 to SB.

> A. Select the menu *Transform→Recode into different variables*.
> B. Select the variable grades and click on the arrow button in the dialog box to move it to Input Variable.
> C. Enter the name of the new variable in the *Name* field of the dialog box and click the *Change* button.
> D. Click the *Old and New Values* button and enter the recoding rules, then click the *Continue* and *OK* buttons.

iii. Calculate the mode and the median.

> A. Select the menu *Analyze→Descriptive Statistics→Frequencies*.
> B. Select the recoded variable in the *Variables* field of the dialog box.
> C. Click the *Statistics* button, select the requested statistics, and click the *Continue* and *OK* buttons.

(c) To conduct a study on the height of university students, we selected, through a random sampling process, a sample of 30 students, obtaining the following results (measured in centimeters):

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.

i. Create the variable height and enter the data.

ii. Obtain a summary of statistics showing the arithmetic mean, median, mode, variance, standard deviation, and quartiles. Interpret the statistics.

> A. Select the menu *Analyze→Descriptive Statistics→Frequencies.*
> B. Select the variable height in the *Variables* field of the dialog box.
> C. Click the *Statistics* button, select the requested statistics, and click the *Continue* and *OK* buttons.

iii. Calculate the third decile and interpret it.

> Follow the same steps as in the previous sections, activating the *Percentiles* option and entering the desired percentile in the corresponding text box.

iv. With the data obtained in the previous sections, calculate the Pearson variation coefficient and the interquartile range, and interpret the results.

(d) To conduct a study on the height of university students, we selected, through a random sampling process, a sample of 30 students, obtaining the following results (measured in centimeters):

x_i	Mark	n_i	f_i	N_i	F_i
[150,160)	155	2	0.07	2	0.07
[160,170)	165	7	0.23	9	0.3
[170,180)	175	12	0.4	21	0.7
[180,190)	185	7	0.23	28	0.93
[190,200)	195	2	0.07	30	1

i. Create the variable height, in which we will enter the class marks, and create the variable frequencies, in which the absolute frequencies will be entered.

ii. Weight the cases of the variable height with the frequencies of the variable frequencies

> A. Select the menu *Data→Weight Cases.*
> B. Activate the *Weight cases by* option, select the variable frequencies, and click the *OK* button.

iii. Obtain a summary of statistics showing the arithmetic mean, median, mode, variance, standard deviation, and quartiles.

> A. Select the menu *Analyze→Descriptive Statistics→Frequencies.*
> B. Select the variable height in the *Variables* field of the dialog box.
> C. Click the *Statistics* button, select the requested statistics, and click the *Continue* and *OK* buttons.

Are there differences between these statistics and those from the previous exercise? What are they due to?

iv. Calculate the third decile.

> Follow the same steps as in the previous sections, activating the *Percentiles* option and entering the corresponding percentile in the text box.

v. Calculate the 62nd percentile.

> Follow the steps from the previous sections, selecting the desired statistic.

vi. With the data obtained in the previous sections, calculate the Pearson variation coefficient and the interquartile range, and interpret the results.

(e) In a hospital, the concentration of immunoglobulin M antibodies in the blood serum of healthy individuals was recorded, resulting in the following data per liter. The gender of the person is indicated in parentheses (H for male and M for female).

(H) 1.071	(H) 0.955	(H) 0.73	(M) 0.908	(M) 0.859
(H) 0.927	(M) 0.962	(M) 1.543	(H) 1.094	(M) 0.847
(H) 1.214	(M) 1.456	(M) 1.516	(M) 1.002	(M) 0.799
(M) 0.881	(M) 1.096	(M) 0.964	(H) 0.973	(H) 1.222
(H) 0.887	(H) 1.022	(M) 0.881	(M) 1.42	(M) 1.205

i. Create the variables gender and immunoglobulin and enter the data.

ii. Split the file, using the variable gender as the segmentation variable

> A. Select the menu *Data→Split File...*
>
> B. Select the option *Compare groups* or *Organize output by groups* (They differ in the way results are presented).
>
> C. Select the variable gender in the *Groups Based on* field of the dialog box and click the *OK* button.

iii. Calculate the arithmetic mean, mode, and median of immunoglobulin, both in men and women.

> A. Select the menu *Analyze→Descriptive Statistics→Frequencies*.
>
> B. Select the variable immunoglobulin in the *Variables* field of the dialog box.
>
> C. Click the *Statistics* button, select the requested statistics, click the *Continue* button, and finally click the *OK* button.

iv. Calculate the variance and standard deviation both in men and women.

> Follow the same steps as the previous section.

v. In which population is the mean more representative, in men or in women?

> To answer this question, it will be necessary to calculate the coefficient of variation.

(f) Two batches of a certain drug are received, manufactured with two different machinery models, one from Madrid and the other from Valencia. A sample of 10 boxes is taken, five from each batch, and the concentration of the active ingredient is measured, obtaining the following results:

Machinery Model	A	B	A	B	A
Origin	Madrid	Madrid	Valencia	Madrid	Valencia
Concentration (mg/mm^3)	17.6	19.2	21.3	15.1	17.6

Machinery Model	B	A	B	B	A
Origin	Valencia	Madrid	Valencia	Madrid	Valencia
Concentration (mg/mm^3)	18.9	16.2	18.3	19	16.4

i. Create the variables origin, machinery and concentration and enter the data.

ii. Calculate the arithmetic mean, standard deviation, skewness coefficient, and kurtosis of the concentration according to the place of origin.

> A. Select the menu *Data→Split File...*
> B. Select the option *Compare groups* or *Organize output by groups*.
> C. Select the variable origin in the *Groups Based on* field of the dialog box and click the *OK* button.
> D. Follow the same steps as the previous exercise to select the statistics.

iii. Draw the box plot of the active ingredient concentration, according to the manufacturing machinery.

> A. Select the menu *Data→Split File...*
> B. Select the option *Compare groups* or *Organize output by groups*.
> C. Select the variable machinery in the *Groups Based on* field of the dialog box and click the *OK* button.
> D. Select the menu *Graphs→Legacy Dialogs→Boxplot...*.
> E. Select the option *Summaries for groups of cases* and click the *Define* button.
> F. Select the variable concentration in the *Boxes Represent* field of the dialog box and click the *OK* button.

3.3 Proposed Exercises

(a) The number of injuries suffered during a season by each player of a football team was as follows:

$$0 - 1 - 2 - 1 - 3 - 0 - 1 - 0 - 1 - 2 - 0 - 1$$
$$1 - 1 - 2 - 0 - 1 - 3 - 2 - 1 - 2 - 1 - 0 - 1$$

i. Create the variable injuries and enter the data. If the data is already available, simply retrieve it.
ii. Calculate: arithmetic mean, median, mode, variance, and standard deviation.
iii. Calculate the skewness and kurtosis coefficients and interpret the results.
iv. Calculate the fourth and eighth deciles.

(b) In a survey on voting intention in an election with three parties A, B, and C, 30 people were asked and the following responses were obtained:

$$A - B - VB - A - C - A - VB - C - A - A - B - B - A - B - B$$
$$B - A - A - C - B - B - B - A - VB - A - B - VB - A - B - B$$

i. Create the variable vote and enter the data.
ii. Calculate the possible statistics for this attribute.

(c) The following table shows the distribution of scores obtained by a group of students.

0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
7	8	13	6	7	6	6	5	6	2

i. Calculate the arithmetic mean, median, and mode.
ii. Calculate the 92nd percentile.
iii. Calculate the standard deviation.
iv. Calculate the skewness coefficient.
v. Calculate the kurtosis coefficient.

(d) In a population study, a sample of 27 people was taken, and they were asked about their age and marital status, obtaining the following results:

i. Create the appropriate variables and enter the data.

Marital Status	Married	Single	Single	Widowed	Married	Married	Divorced	Single	Single	Single	Widowed	Married	Single	Divorced	Widowed	Divorced	Single	Widowed	Married	Widowed	Married	Divorced	Divorced	Widowed	Widowed	Single	Widowed
Age	62	31	45	100	39	62	31	45	100	21	38	59	62	65	38	59	62	65	21	31	62	59	65	38	59	31	65

ii. Calculate the mean and standard deviation of age according to marital status.

iii. Draw the bar chart for the absolute frequencies of age according to marital status.

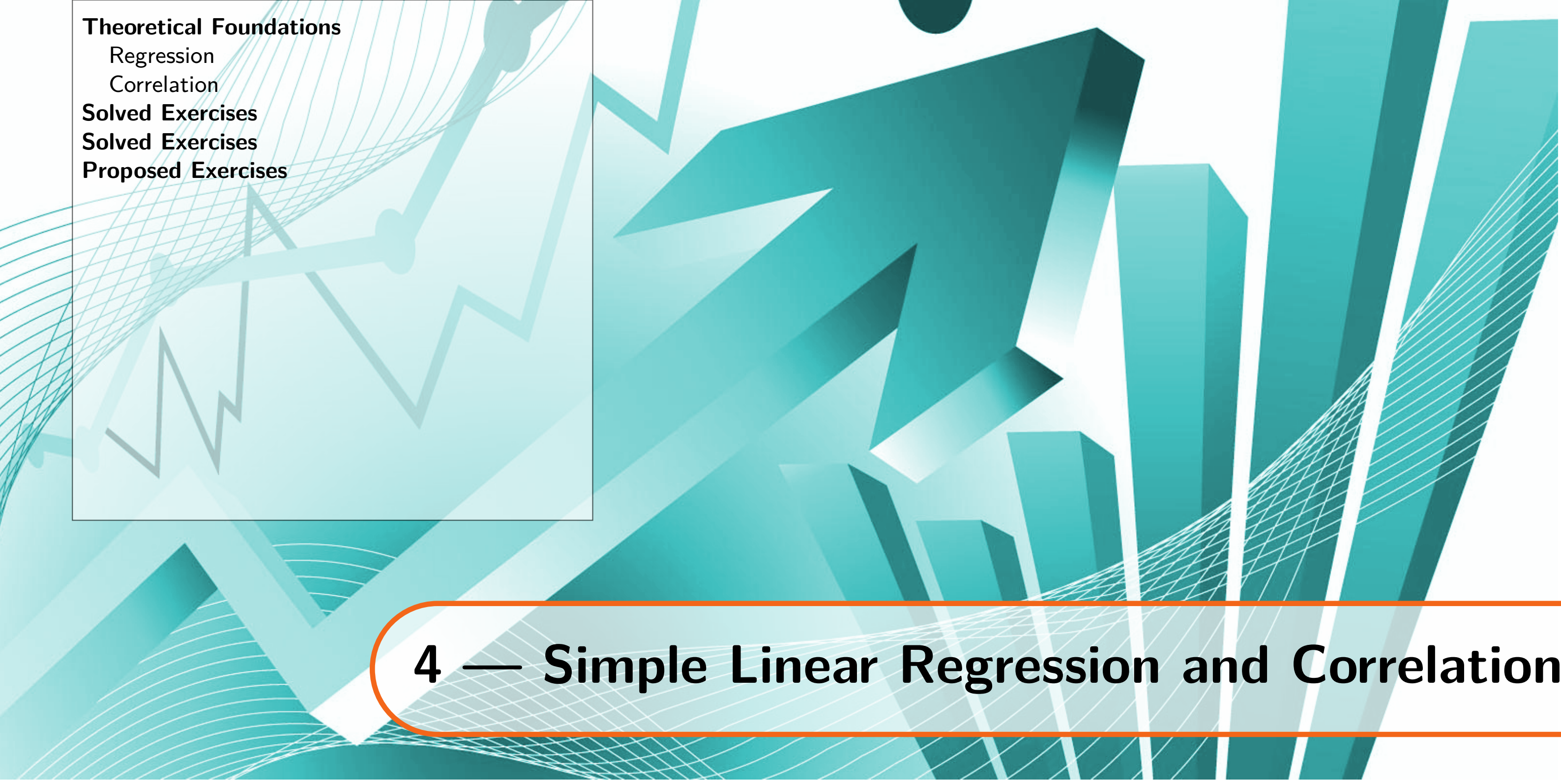
(e) In a study, the blood pressure of 25 individuals was measured. They were also asked if they smoke and drink:

Smoker	yes	no	yes	yes	yes	no	no	yes	no	yes	no	yes	no	yes	no	no	yes	no	no	no	yes	no	yes	no	yes
Drinker	no	no	yes	yes	no	no	yes	yes	no	yes	no	yes	yes	yes	no	yes	yes	no	no	yes	yes	yes	no	yes	no
Blood Pressure	80	92	75	56	89	93	101	67	89	63	98	58	91	71	52	98	104	57	89	70	93	69	82	70	49

i. Create the corresponding variables and enter the data.

ii. Calculate the arithmetic mean, standard deviation, skewness coefficient, and kurtosis of blood pressure by groups depending on whether they drink and/or smoke.

iii. Draw the histogram for the absolute frequencies of blood pressure according to the groups created earlier.



4 — Simple Linear Regression and Correlation

4.1 Theoretical Foundations

4.1.1 Regression

Regression is the part of statistics that deals with determining the possible relationship between a numerical variable Y, usually called the *dependent variable*, and another set of numerical variables, $X_1, X_2, \ldots, X_n$, known as *independent variables*, of the same population. This relationship is reflected through a functional model $y = f(x_1, \ldots, x_n)$.

The simplest case occurs when there is only one independent variable X, and then it is called *simple regression.* In this case, the model that explains the relationship between X and Y is a function of one variable $y = f(x)$.

Depending on the form of this function, there are many types of simple regression. The most common ones are listed in the following table:

Family of Curves	Generic Equation
Linear	$y = b_0 + b_1 x$
Quadratic	$y = b_0 + b_1 x + b_2 x^2$
Cubic	$y = b_0 + b_1 x + b_2 x^2 + b_3 x^3$
Power	$y = b_0 \cdot x^{b_1}$
Exponential	$y = b_0 \cdot e^{b_1 x}$
Logarithmic	$y = b_0 + b_1 \ln x$
Inverse	$y = b_0 + \frac{b_1}{x}$
Compound	$y = b_0 b_1^x$
Growth	$y = e^{b_0 + b_1 x}$
G (S-Curve)	$y = e^{b_0 + \frac{b_1}{x}}$

To choose one type of model or another, the *scatter plot* is usually represented, which consists of plotting on Cartesian axes corresponding to the variables X and Y, the pairs of values (x_i, y_j) observed in each individual of the sample.

■ **Example 4.1** In figure 4.1, the scatter plot corresponding to a sample of 30 individuals in which height in cm (X) and weight in kg (Y) have been measured is shown. In this case, the shape of the point cloud reflects a linear relationship between height and weight. ■

Depending on the shape of the point cloud in the diagram, the most appropriate model is chosen (figure 4.2), and the parameters of that model are determined so that the resulting function fits the point cloud as closely as possible.

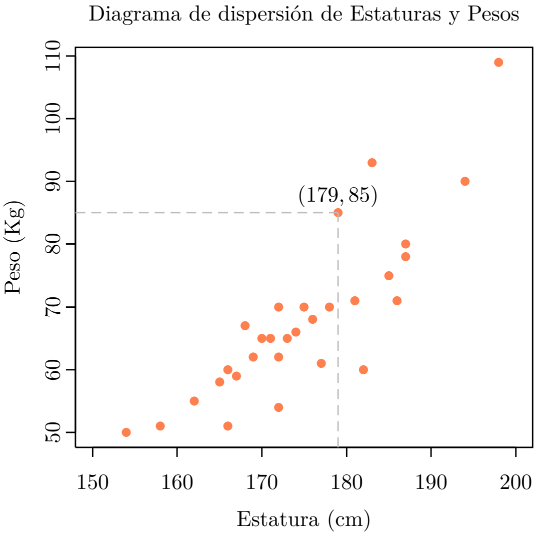
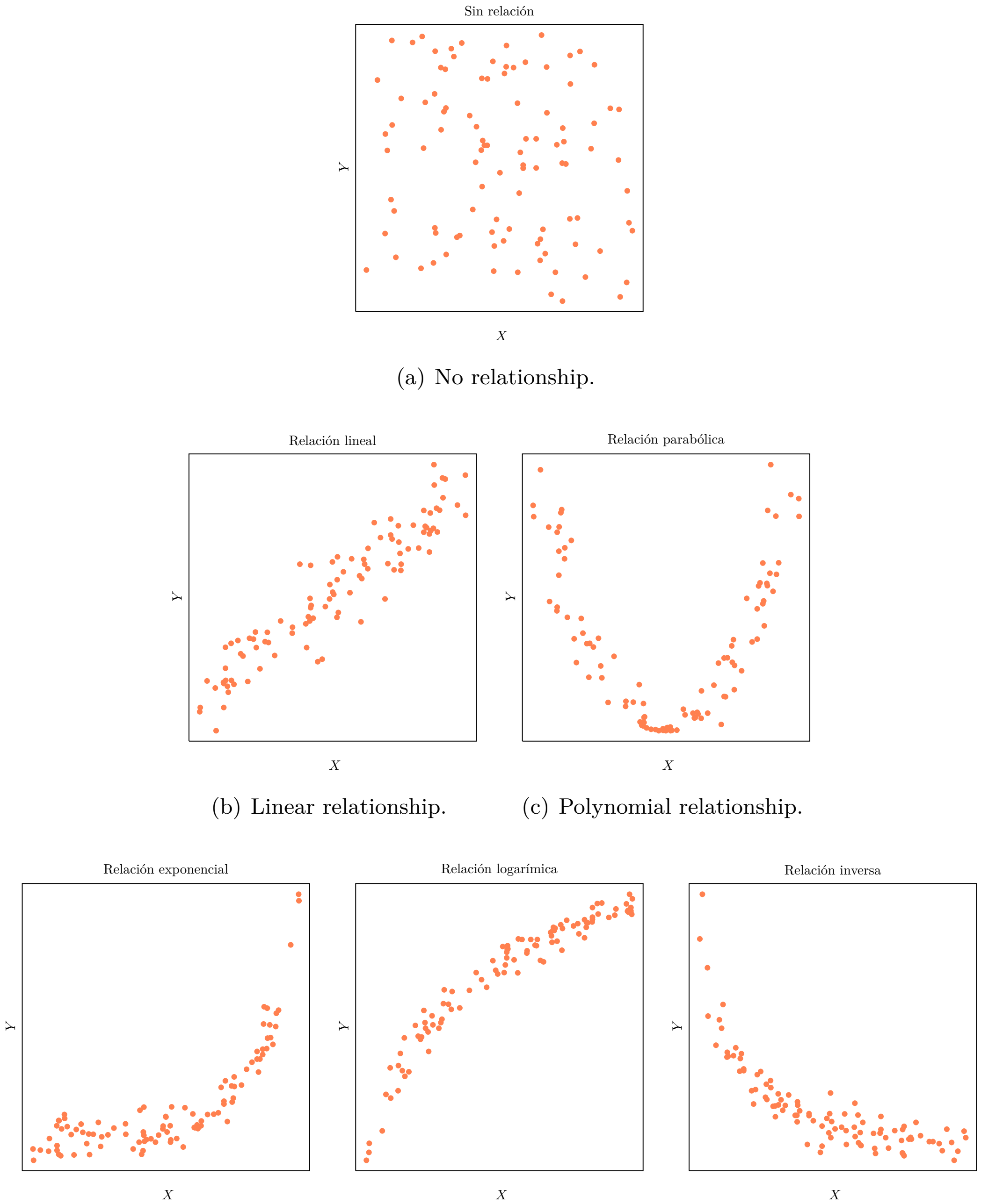


Figure 4.1: Scatter plot. The point (179,85) indicated corresponds to an individual in the sample who is 179 cm tall and weighs 85 kg.



(a) No relationship.

(b) Linear relationship.

(c) Polynomial relationship.

(d) Exponential relationship.

(e) Logarithmic relationship.

(f) Inverse relationship.

Figure 4.2: Scatter plots corresponding to different types of relationships between variables.

The criterion usually used to obtain the optimal function is that the distance of each point to the curve, measured on the Y-axis, is as small as possible. These distances are called *residuals* or *errors* in Y (figure 4.3). The function that best fits the point cloud will, therefore, be the one that minimizes the sum of the squares of the residuals.*

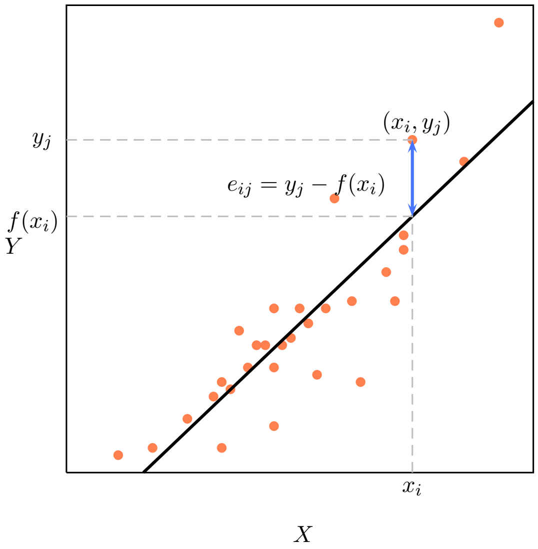


Figure 4.3: Residuals or errors in Y. The residual corresponding to a point (x_i, y_j) is the difference between the value y_j observed in the sample and the theoretical value of the model $f(x_i)$, that is, $e_{ij} = y_j - f(x_i)$.

Regression Lines

In the case that the point cloud has a linear shape and we choose to explain the relationship between X and Y using a line $y = a + bx$, the parameters to be determined are a (intercept with the y-axis) and b (slope of the line). The values of these parameters that minimize the sum of squared residuals determine the optimal line. This line is known as the *regression line of Y on X* and explains the variable Y as a function of the variable X. Its equation is

$$y = \bar{y} + \frac{s_{xy}}{s_x^2}(x - \bar{x}),$$

where s_{xy} is a statistic called *covariance* that measures the degree of linear relationship, and its formula is

$$s_{xy} = \frac{1}{n}\sum_{i,j}(x_i - \bar{x})(y_j - \bar{y})n_{ij}.$$

■ **Example 4.2** In figure 4.4, the regression lines of Height on Weight and Weight on Height from the previous example are shown. ■

The slope of the regression line of Y on X is known as the *regression coefficient of Y on X*, and it measures the increase that the variable Y will experience for each unit that the variable X increases, according to the line.

The smaller the residuals, in absolute value, the better the model fits the point cloud, and therefore, the better it explains the relationship between X and Y. When all residuals are zero, the line passes through all points in the cloud, and the relationship is perfect. In this case, both lines, the one of Y on X and the one of X on Y, coincide (figure 4.5(a)).

*They are squared to avoid positive and negative residuals canceling each other out in the sum.

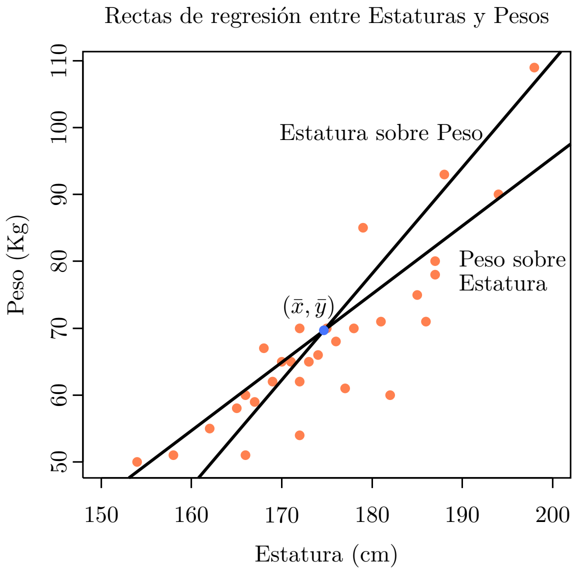


Figure 4.4: Regression lines of Height on Weight and Weight on Height. The regression lines always intersect at the mean point $(\bar{x}, \bar{y})$.

On the other hand, when there is no linear relationship between the variables, the regression line of Y on X has a zero slope, and therefore the equation is $y = \bar{y}$, in which x does not appear, or $x = \bar{x}$ in the case of the regression line X on Y, so that both lines intersect perpendicularly (figure 4.5(b)).

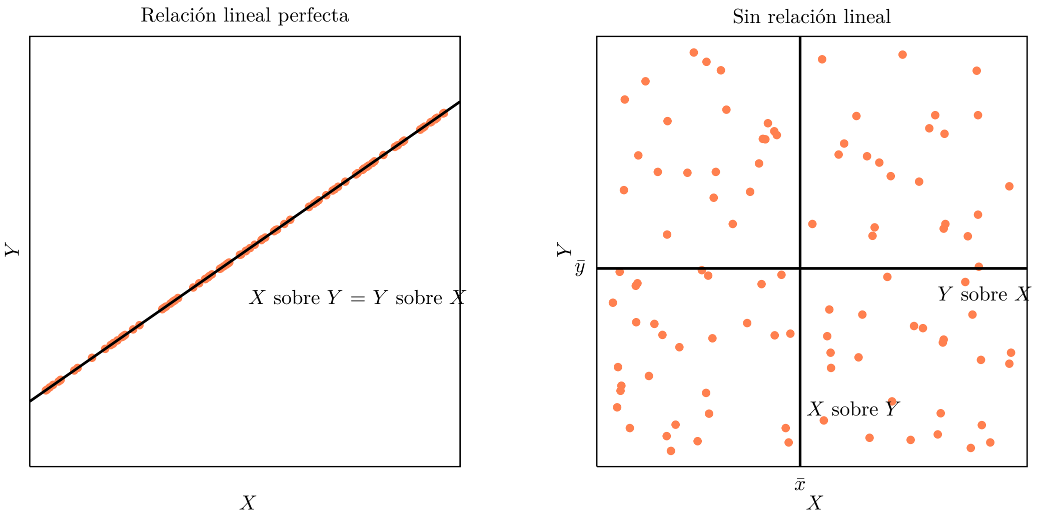
4.1.2 Correlation

The main objective of simple regression is to construct a functional model $y = f(x)$ that best explains the relationship between two variables X (independent variable) and Y (dependent variable) measured in the same sample. Generally, the constructed model is used to make predictive inferences of Y based on X in the rest of the population. However, although regression guarantees that the constructed model is the best possible within the chosen type of model (linear, polynomial, exponential, logarithmic, etc.), it may still not be a good model for making predictions, precisely because there may not be that type of relationship between X and Y. Therefore, to validate a model for making reliable predictions, measures are needed that tell us about the degree of dependence between X and Y with respect to a constructed regression model. These measures are known as *correlation* measures.

Depending on the type of adjusted model, there will be different types of correlation measures. Thus, if the constructed regression model is a line, we will talk about linear correlation; if it is a polynomial, we will talk about polynomial correlation; if it is an exponential function, we will talk about exponential correlation, etc. In any case, these measures will tell us how good the constructed model is, and consequently, whether we can trust the predictions made with that model.

Most correlation measures arise from the study of residuals or errors in Y, which are the distances of the points in the scatter plot to the constructed regression curve, measured on the Y axis, as shown in figure (4.3). These distances are, in reality, the predictive errors of the model on the actual values of the sample.

The smaller the residuals, the better the model fits the point cloud, and therefore, the better it explains the relationship between X and Y. When all residuals are zero, the regression curve passes through all points in the cloud, and then it is said that the relationship is perfect, or that there is a functional dependence between X and Y (figure 4.5(a)). On the other hand,



(a) Linear functional dependence.

(b) Linear independence.

Figure 4.5: Different degrees of dependence. In the first case, the relationship is perfect and the residuals are zero. In the second case, there is no linear relationship and the slope of the line is zero.

when the residuals are large, the model will not explain the relationship between X and Y well, and therefore, its predictions will not be reliable (figure 4.5(b)).

Residual Variance

A first correlation measure, constructed from the residuals, is the *residual variance*, which is defined as the average of the squared residuals:

$$s_{ry}^2 = \frac{\sum_{i,j} e_{ij}^2 n_{ij}}{n} = \frac{\sum_{i,j}(y_j - f(x_i))^2 n_{ij}}{n}.$$

When the residuals are zero, then $s_{ry}^2 = 0$ and that indicates that there is functional dependence. On the other hand, when the variables are independent with respect to the adjusted regression model, then the residuals become the deviations of the Y values with respect to their mean, and it holds that $s_{ry}^2 = s_y^2$. Thus, it holds that

$$0 \leq s_{ry}^2 \leq s_y^2.$$

According to this, the smaller the residual variance, the greater the dependence between X and Y, according to the adjusted model. However, the variance has units of Y squared, which makes its interpretation difficult.

Coefficient of Determination

Since the maximum value that the residual variance can take is the variance of Y, a coefficient can be easily defined from the comparison of both measures. Thus arises the *coefficient of determination*, which is defined as

$$R^2 = 1 - \frac{s_{ry}^2}{s_y^2}.$$

It holds that

$$0 \leq R^2 \leq 1,$$

and it also has no units, making it easier to interpret than the residual variance:

- $R^2 = 0$ indicates independence according to the type of relationship proposed by the regression model.
- $R^2 = 1$ indicates functional dependence.

Therefore, the higher R^2 is, the better the regression model.

If we multiply the coefficient of determination by 100, we obtain the percentage of variability of Y explained by the regression model. The remaining percentage corresponds to the variability that remains to be explained and corresponds to the predictive error of the model. Thus, for example, if we have a coefficient of determination $R^2 = 0.5$, the regression model would explain half of the variability of Y, and consequently, if that model is used to make predictions, they would have half the error compared to not using it and taking the mean value of Y as the prediction.

Linear Coefficient of Determination

In the case that the regression model is linear, the formula for the coefficient of determination simplifies and becomes

$$r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2},$$

which is known as the *linear coefficient of determination.*

Correlation Coefficient

Another common measure of dependence is the *correlation coefficient*, which is defined as the square root of the coefficient of determination:

$$R = \pm\sqrt{1 - \frac{s_{ry}^2}{s_y^2}},$$

taking the root with the same sign as the covariance.

The only advantage of the correlation coefficient over the coefficient of determination is that it has a sign, and therefore, in addition to the degree of dependence between X and Y, it also tells us whether the relationship is direct (positive sign) or inverse (negative sign). Its interpretation is:

- $R = 0$ indicates independence with respect to the type of relationship proposed by the regression model.
- $R = -1$ indicates inverse functional dependence.
- $R = 1$ indicates direct functional dependence.

Therefore, the closer it is to -1 or 1, the better the regression model.

Linear Correlation Coefficient As with the coefficient of determination, when the regression model is linear, the formula for the correlation coefficient becomes

$$r = \frac{s_{xy}}{s_x s_y},$$

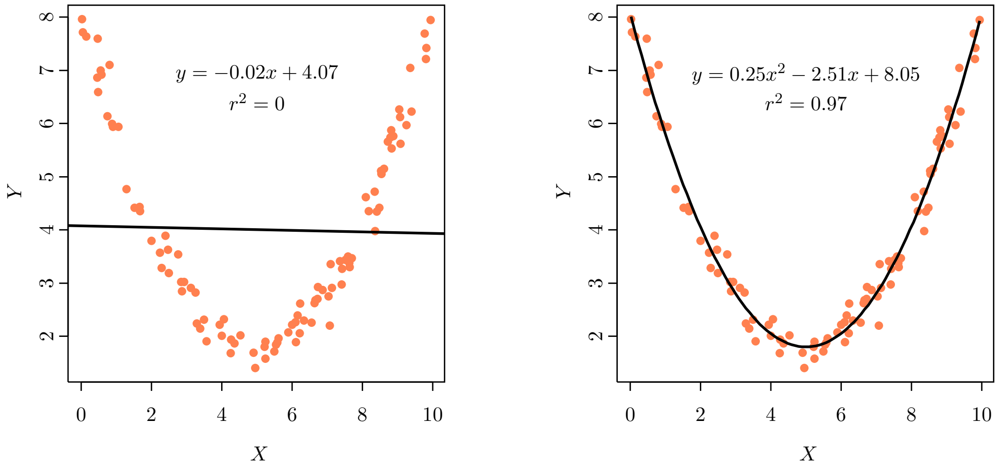
and it is called the *linear correlation coefficient.*

Finally, it is worth noting that a zero coefficient of determination or correlation indicates independence according to the constructed regression model, but there may be dependence of another type. This is clearly seen in the example in figure 4.6.

Reliability of Predictions

Although the coefficient of determination or correlation tells us about the goodness of a regression model, it is not the only data to consider when making predictions.

The reliability of the predictions we make with a regression model depends on several factors:



(a) Weak linear dependence.

(b) Strong parabolic dependence.

Figure 4.6: In the figure on the left, a linear model has been adjusted and an $R^2 = 0$ has been obtained, indicating that the model does not explain the relationship between X and Y at all, but we cannot say that X and Y are independent. In fact, in the figure on the right, it is observed that by adjusting a parabolic model, $R^2 = 0.97$, indicating that there is almost a parabolic functional dependence between X and Y.

- The coefficient of determination: The higher it is, the smaller the predictive errors and the greater the reliability of the predictions.
- The variability of the population: The more variable a population is, the more difficult it is to predict, and therefore, the less reliable the model's predictions will be.
- The sample size: The larger it is, the more information we have, and consequently, the more reliable the predictions will be.

Additionally, it should be noted that a regression model is valid for the range of values observed in the sample, but outside that range, we have no information about the type of relationship between the variables, so we should not make predictions for values that are far from those observed in the sample.

4.2 Solved Exercises

(a) Two variables A and B have been measured in 10 individuals, obtaining the following results:

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

i. Create the variables A and B and enter these data.

ii. Draw the corresponding scatter plot.

> A. Select the menu *Graphs→Legacy Dialogs→Scatter/Dot...*, choose the option *Simple Scatter* and click the *Define* button.
> B. Select the variable `B` in the *Y Axis* field of the dialog box.
> C. Select the variable `A` in the *X Axis* field of the dialog box and click the *OK* button.

Based on the diagram, what type of model do you think will best explain the relationship between B and A?

iii. Calculate the regression line of B on A.

> A. Select the menu *Analyze→Regression→Linear...*.
> B. Select the variable `B` in the *Dependent* field of the dialog box.
> C. Select the variable `A` in the *Independent* field of the dialog box and click the *OK* button.
> D. To write the equation of the line, observe the table named Coefficients in the results window, and in the B column of the Unstandardized Coefficients, find the constant of the line in the first row and the slope in the second row.

iv. Draw this line on the scatter plot.

> A. Edit the previously created chart by double-clicking on it.
> B. Select the points by clicking on one of them.
> C. Select the menu *Elements→Fit Line at Total* (You can also use the toolbar instead of the menu)
> D. Close the Properties window.
> E. Close the chart editor by closing the window.

v. Calculate the regression line of A on B and draw it on the corresponding scatter plot.

> Repeat the steps from the previous sections but choose the variable *Dependent* as `A`, and the variable *Independent* as `B`.

vi. Are the residuals large? Comment on the results.

(b) In a degree program, the relationship between the average number of daily study hours and the number of failed subjects is to be studied. The following sample was obtained:

Hours	Fails	Hours	Fails	Hours	Fails
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

i. Create the variables hours and fails and enter these data.

ii. Calculate the regression line of fails on hours and draw it.

> A. Select the menu *Analyze→Regression→Linear...*.
> B. Select the variable fails in the *Dependent* field of the dialog box.
> C. Select the variable *hours* in the *Independent* field of the dialog box and click the *OK* button.
> D. To write the equation of the line, observe the table named Coefficients in the results window, and in the B column of the Unstandardized Coefficients, find the constant of the line in the first row and the slope in the second row.
> E. Select the menu *Graphs→Legacy Dialogs→Scatter/Dot...*, choose the option *Simple Scatter* and click the *Define* button.
> F. Select the variable fails in the *Y Axis* field of the dialog box.
> G. Select the variable hours in the *X Axis* field of the dialog box and click the *OK* button.
> H. Edit the created chart by double-clicking on it.
> I. Select the points by clicking on one of them.
> J. Select the menu *Elements→Fit Line at Total* (You can also use the toolbar instead of the menu)
> K. Close the Properties window.
> L. Close the chart editor by closing the window.

iii. Indicate the regression coefficient of fails on hours. How would you interpret it?

> The regression coefficient is the slope of the regression line, which in this case is -1.23 and indicates that for each additional hour of study, there are 1.23 fewer fails.

iv. Is the linear relationship between these two variables better or worse than in the previous exercise? Comment on the results based on the regression line charts and their residuals.

> The linear relationship between these two variables is worse than in the previous exercise, as there are residuals in this case.

v. Calculate the linear correlation and determination coefficients. Is the regression line a good model? What percentage of the variability in the number of fails is explained by the model?

> Observe the table named Model Summary in the results window, where the values of the linear correlation coefficient R and the linear determination coefficient R squared are found.

vi. Use the regression line to predict the number of fails corresponding to 3 hours of daily study. Is this prediction reliable?

> A. Create a new variable values and enter the study hours for which we want to predict.
> B. Select the menu *Transform→Compute Variable...*.
> C. Enter the name of the new variable prediction in the *Target Variable* field of the dialog box.
> D. Enter the equation of the line in the *Numeric Expression* field, using the previously calculated coefficients and the variable values and click the *OK* button.

vii. According to the linear model, how many daily study hours will a student need to pass everything?

> Follow the same steps as in the previous sections, but choose the dependent variable hours, and the independent variable fails.

4.3 Solved Exercises

i. After drinking a liter of wine, the blood alcohol concentration was measured at different times, obtaining:

Time after (minutes)	30	60	90	120	150	180	210
Concentration (grams/liter)	1.6	1.7	1.5	1.1	0.7	0.2	2.1

A. Create the variables time and alcohol and enter these data.

B. Calculate the linear correlation coefficient and interpret it.

> Select the menu *Analyze→Correlate→Bivariate...*.
>
> Select both variables in the *Variables* field of the dialog box and click the *OK* button.

C. Draw the scatter plot along with the fitted line corresponding to alcohol on time. Is there any individual with a residual that is too large? If so, remove that individual from the sample and recalculate the correlation coefficient. Has the model improved?

> Select the menu *Graphs→Legacy Dialogs→Scatter/Dot...*, choose the option *Simple Scatter* and click the *Define* button.
>
> Select the variable alcohol in the *Y Axis* field of the dialog box.
>
> Select the variable time in the *X Axis* field of the dialog box and click the *OK* button.
>
> Edit the previously created chart by double-clicking on it.
>
> Select the points by clicking on one of them.
>
> Select the menu *Elements→Fit Line at Total* (You can also use the toolbar instead of the menu)
>
> Close the Properties window.

> Close the chart editor by closing the window.
>
> If there is any individual with a residual that is too large, go to the *Data Editor* window and remove them.
>
> Repeat the steps from the previous section.

D. If the maximum blood alcohol concentration allowed by law for driving is 0.5 g/l, how long will you have to wait after drinking a liter of wine to be able to drive without breaking the law? Is this prediction reliable?

> i
>
> Select the menu *Analyze→Regression→Linear...*.
>
> Select the variable time in the *Dependent* field of the dialog box.
>
> Select the variable alcohol in the *Independent* field of the dialog box and click the *OK* button.
>
> To write the equation of the line, observe the table named Coefficients in the results window, and in the B column of the Unstandardized Coefficients, find the constant of the line in the first row and the slope in the second row.
>
> Create a new variable values and enter the values you want to study.
>
> Select the menu *Transform→Compute Variable...*.
>
> Enter the name of the new variable prediction in the *Target Variable* field of the dialog box.
>
> Enter the equation of the line in the *Numeric Expression* field, using the previously mentioned coefficients and the variable values and click the *OK* button.

4.4 Proposed Exercises

i. The loss of activity experienced by a drug from the moment of its manufacture over time is determined, obtaining the following result:

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

The following is required:

A. The fundamental relationship (regression line) between remaining activity and elapsed time.

B. By what percentage does the activity decrease each year?

C. How long must pass for the drug to have an activity of 80

ii. In a study on the dosage of a certain drug, 6 patients were treated with daily doses of 2 mg, 7 patients with 3 mg, and another 7 patients with 4 mg. Of the patients treated with 2 mg, 2 were cured after 5 days, and 4 after 6 days. Of the patients treated with 3 mg daily, 2 were cured after 3 days, 4 after 5 days, and 1 after 6 days. And of the patients treated with 4 mg daily, 5 were cured after 3 days and 2 after 5 days. The following is required:

A. Calculate the regression line of the healing time with respect to the administered dose.

B. Calculate the regression coefficients. Interpret the results.

C. Determine the expected healing time for a dose of 5 mg daily. Is this prediction reliable?

D. What dose should be applied if we want the patient to take 4 days to heal? Is the prediction reliable?

5 — Nonlinear Regression

5.1 Theoretical Foundations

Simple regression aims to construct a functional model $y = f(x)$ that best explains the relationship between two variables Y (dependent variable) and X (independent variable) measured in the same sample.

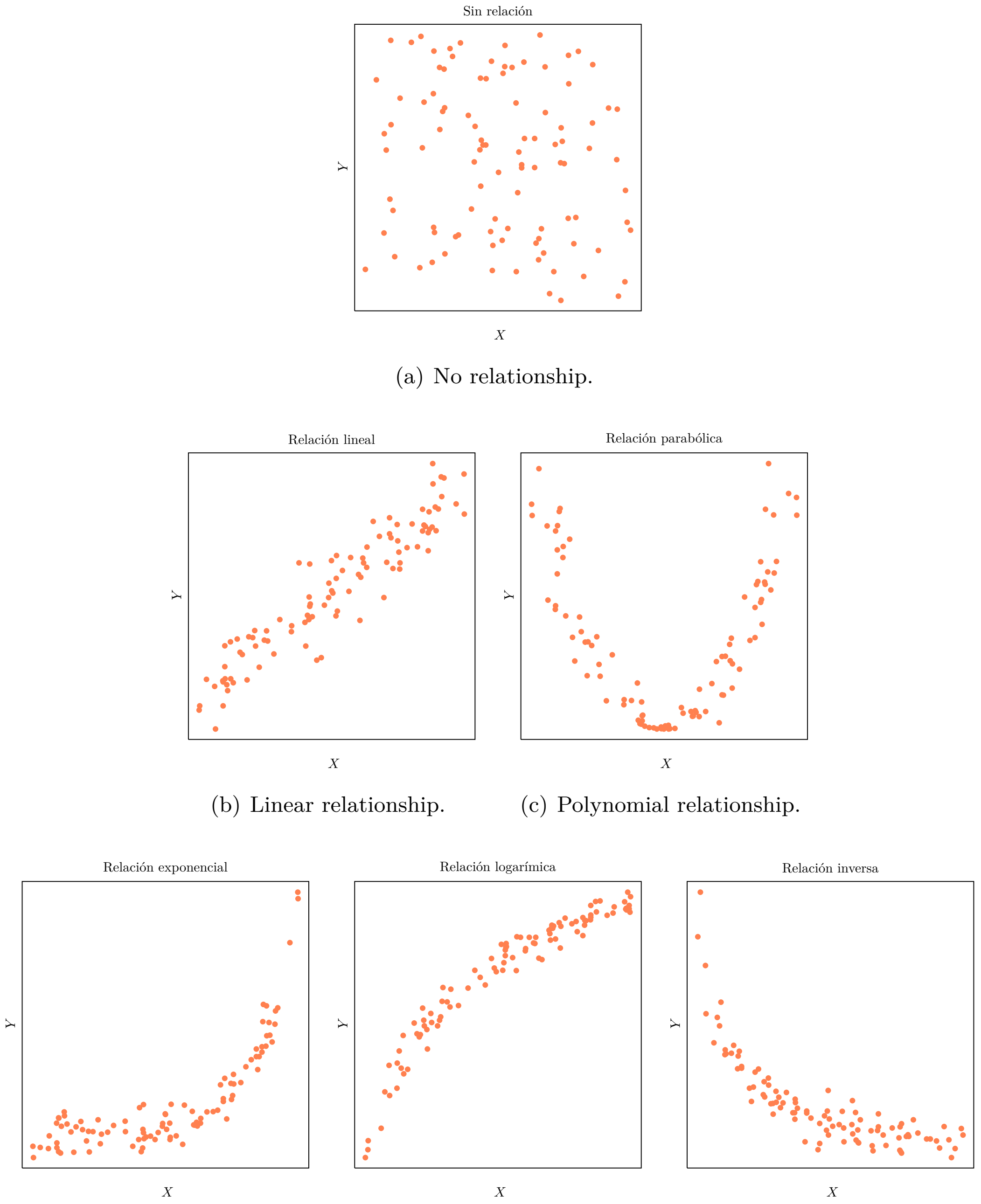
We have already seen that, depending on the form of this function, there are many types of simple regression. Among the most common are:

Family of Curves	Generic Equation
Linear	$y = b_0 + b_1 x$
Quadratic	$y = b_0 + b_1 x + b_2 x^2$
Cubic	$y = b_0 + b_1 x + b_2 x^2 + b_3 x^3$
Power	$y = b_0 \cdot x^{b_1}$
Exponential	$y = b_0 \cdot e^{b_1 x}$
Logarithmic	$y = b_0 + b_1 \ln x$
Inverse	$y = b_0 + \frac{b_1}{x}$
Compound	$y = b_0 b_1^x$
Growth	$y = e^{b_0 + b_1 x}$
G (S-Curve)	$y = e^{b_0 + \frac{b_1}{x}}$

The choice of one type of model or another is usually made according to the shape of the point cloud in the scatter plot. Sometimes it will be clear what type of model should be constructed, as in the scatter plots in figure 5.1. But other times it will not be so clear, and in these cases, it is common to fit the two or three models that seem most convincing, and then choose the one that best explains the relationship between Y and X, looking at the coefficient of determination[*] of each model.

We have already seen in the practice on simple linear regression how to construct regression lines. In the case that we choose to fit a nonlinear model, the construction can be done following the same steps as in the linear case. Basically, it is about determining the parameters of the model that minimize the sum of the squares of the residuals in Y. In the power and exponential models, the system applies logarithmic transformations to the variables and then fits a linear model to the transformed data. In the inverse model, the system replaces the dependent variable with its inverse before estimating the regression equation.

[*] See the correlation practice.



(a) No relationship.

(b) Linear relationship.

(c) Polynomial relationship.

(d) Exponential relationship.

(e) Logarithmic relationship.

(f) Inverse relationship.

Figure 5.1: Scatter plots corresponding to different types of relationships between variables.

5.2 Solved Exercises

i. In an experiment, the number of bacteria per unit volume in a culture was measured every hour, obtaining the following results:

Hours	0	1	2	3	4	5	6	7	8
No. of Bacteria	25	32	47	65	92	132	190	275	362

Tasks:

A. Create the variables hours and bacteria and enter these data.

B. Draw the corresponding scatter plot. Based on the diagram, what type of model do you think will best explain the relationship between the number of bacteria and the elapsed time?

> *i*
>
> Select the menu *Graphs→Legacy Dialogs→Scatter/Dot...*, choose the option *Simple Scatter* and click the *Define* button.
>
> Select the variable bacteria in the *Y Axis* field of the dialog box.
>
> Select the variable hours in the *X Axis* field of the dialog box and click the *OK* button.

C. Compare the different regression models based on the coefficient of determination. Which type of model is the best?

> *i*
>
> Select the menu *Analyze→Regression→Curve Estimation...*.
>
> Select the variable bacteria in the *Dependent* field of the dialog box.
>
> Select the variable hours in the *Independent* field of the dialog box.
>
> Uncheck the option *Display models*.
>
> Check the options linear, quadratic, cubic, exponential, and logarithmic, and click the *OK* button.

D. Based on the above, calculate the regression model that best explains the relationship between bacteria and hours.

> *i*
>
> Use the coefficients that appear in the previous point and the table in the theoretical foundations section.

E. According to the previous model, how many bacteria will there be after 3 and a half hours from the start of the culture? And after 10 hours? Are these predictions reliable?

> *i*
>
> Create a new variable values and enter the values of the hours for which we want to predict the bacteria.
>
> Select the menu *Transform→Compute Variable...*.
>
> Enter the name of the new variable prediction in the *Target Variable* field of the dialog box.
>
> Enter the equation of the best model in the *Numeric Expression* field, using the previously obtained coefficients and the variable values and click the *OK* button.

F. Provide the most reliable possible prediction of the time that would need to pass for there to be 100 bacteria in the culture.

> Repeat the steps from the previous section, entering the variable hours in the *Dependent* field and the variable bacteria in the *Independent* field.

ii. Two variables S and T have been measured in 10 individuals, obtaining the following results:

$$(-1.5\,,\,2.25)\,,\ (0.8\,,\,0.64)\,,\ (-0.2\,,\,0.04)\,,\ (-0.8\,,\,0.64)\,,\ (0.4\,,\,0.16)\,,$$
$$(0.2\,,\,0.04)\,,\ (-2.1\,,\,4.41)\,,\ (-0.4\,,\,0.16)\,,\ (1.5\,,\,2.25)\,,\ (2.1\,,\,4.41).$$

A. Create the variables S and T and enter these data.

B. Calculate the regression line of T on S. Draw this line on the scatter plot. Can we say that S and T are independent?

> Select the menu *Analyze→Regression→Linear...*.
>
> Select the variable T in the *Dependent* field of the dialog box.
>
> Select the variable S in the *Independent* field of the dialog box and click the *OK* button.
>
> To write the equation of the line, observe the table named Coefficients in the results window, and in the B column of the Unstandardized Coefficients, find the constant of the line in the first row and the slope in the second row.
>
> Select the menu *Graphs→Legacy Dialogs→Scatter/Dot...*, choose the option *Simple Scatter* and click the *Define* button.
>
> Select the variable T in the *Y Axis* field of the dialog box.
>
> Select the variable S in the *X Axis* field of the dialog box and click the *OK* button.
>
> Edit the previously created chart by double-clicking on it.
>
> Select the points by clicking on one of them.
>
> Select the menu *Elements→Fit Line at Total* (You can also use the toolbar instead of the menu)
>
> Close the Properties window.
>
> Close the chart editor by closing the window.

C. Compare the different regression models based on the coefficient of determination. What type of relationship exists between T and S?

> Select the menu *Analyze→Regression→Curve Estimation...*.
>
> Select the variable T in the *Dependent* field of the dialog box.
>
> Select the variable S in the *Independent* field of the dialog box.
>
> Uncheck the option *Display models*.
>
> Check the options linear, quadratic, cubic, and exponential and click the *OK* button.

D. Based on the above, fit the most appropriate regression model.

> Use the coefficients that appear in the previous point and the table in the theoretical foundations section.

5.3 Proposed Exercises

i. In a dietary center, a new weight loss diet is being tested on a sample of 12 individuals. For each of them, the number of days they have been on the diet and the number of kilograms lost since then have been measured, obtaining the following results:

$$(33\ ,\ 3.9),\ (51\ ,\ 5.9),\ (30\ ,\ 3.2),\ (55\ ,\ 6.0),\ (38\ ,\ 4.9),\ (62\ ,\ 6.2),$$
$$(35\ ,\ 4.5),\ (60\ ,\ 6.1),\ (44\ ,\ 5.6),\ (69\ ,\ 6.2),\ (47\ ,\ 5.8),\ (40\ ,\ 5.3)$$

A. Draw the scatter plot. Based on the point cloud, what type of model would best explain the relationship between the kilograms lost and the days on the diet?

B. Construct the regression model that best explains the relationship between the kilograms lost and the days on the diet.

C. Use the constructed model to predict the number of kilograms lost after 40 days on the diet and after 100 days. Are these predictions reliable?

ii. The concentration of a drug in the blood, C in mg/dl, is a function of time, t in hours, and is given by the following table:

t	2	3	4	5	6	7	8
C	25	36	48	64	86	114	168

A. According to the exponential model, what would be the drug concentration at 4.8 hours? Is the prediction reliable? Justify your answer adequately.

B. According to the logarithmic model, how much time must pass for the concentration to be 100 mg/dl?

6 — Confidence Intervals for Means and Proportions

6.1 Theoretical Foundations

6.1.1 Statistical Inference and Parameter Estimation

The objective of a statistical study is twofold: to describe the sample chosen from a population in which some characteristic is to be studied, and to make inferences, that is, to draw conclusions about the population from which the sample was taken.

The methodology that leads to conclusions about the population, based on the information contained in the sample, constitutes *Statistical Inference*.

Since the sample contains less information than the population, the conclusions about the population will be approximate. Therefore, one of the objectives of statistical inference is to determine the probability that a conclusion obtained from the analysis of a sample is true, and for this, it relies on probability theory.

When the value of one of the population parameters is to be known, the procedure to be used is *Parameter Estimation*, which in turn is divided into *Point Estimation*, when a single value is given as an estimate of the population parameter considered, and *Interval Estimation*, when it is of interest to know not only an approximate value of the parameter but also the precision of the estimate. In the latter case, the result is an interval, within which the true value of the population parameter will be, with a certain confidence. This interval is called a *confidence interval*. Unlike point estimation, in which a single estimator is used, interval estimation uses two estimators, one for each end of the interval.

6.1.2 Confidence Intervals

Given two sample statistics L_1 and L_2, the interval $I = (L_1,\ L_2)$ is said to be a *Confidence Interval* for a population parameter θ, with *confidence level* $1-\alpha$ (or *significance level* α), if the probability that the statistics that determine the limits of the interval take values such that θ is between them is equal to $1-\alpha$, that is,

$$P\left(L_1 < \theta < L_2\right) = 1 - \alpha$$

The ends of the interval are random variables whose values depend on the sample considered. That is, the lower and upper ends of the interval would be $L_1\left(X_1, ..., X_n\right)$ and $L_2\left(X_1, ..., X_n\right)$ respectively, although we will usually write L_1 and L_2 to simplify notation. We will denote by l_1 and l_2 the values taken by these variables for a given sample $\left(x_1, ..., x_n\right)$.

When the definition says that the probability that the parameter θ is in the interval $(L_1,\ L_2)$ is $1-\alpha$, it means that in $100\,(1-\alpha)\ \%$ of the possible samples, the value of θ would be in the corresponding intervals $(l_1,\ l_2)$.

Once a sample is taken, and from it the corresponding interval $(l_1,\ l_2)$ is determined, it would not make sense to talk about the probability that the parameter θ is in the interval $(l_1,\ l_2)$, since l_1 and l_2 are numbers, the parameter θ, which is also a number, although unknown, will either be in that interval or not, and therefore we talk about confidence instead of probability.

Thus, when we talk about a confidence interval for the parameter θ with a confidence level of $1-\alpha$, we understand that before taking a sample, there is a probability of $1-\alpha$ that the interval constructed from it will contain the value of the parameter θ. Or, in other words, if we took all possible samples of the same size and calculated their respective intervals, $100(1-\alpha)\%$ of these would contain the true value of the parameter to be estimated (see figure 6.1).

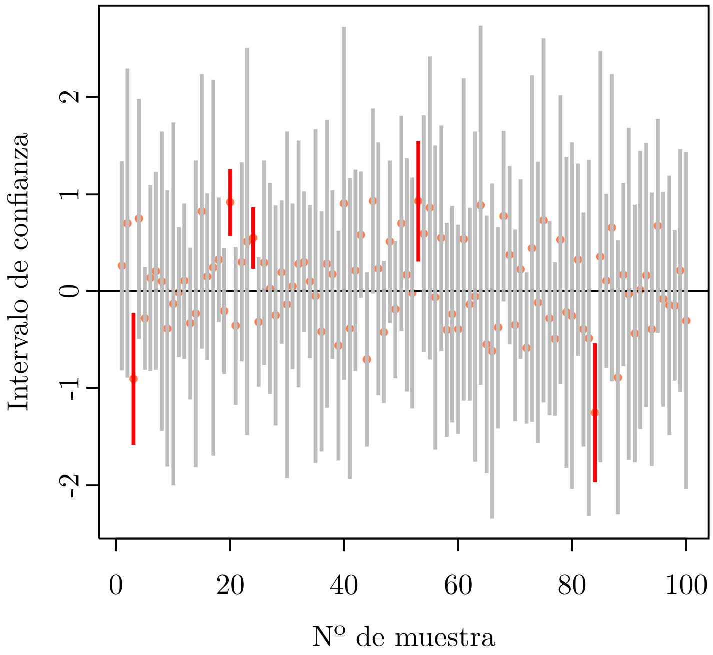


Figure 6.1: 95% confidence intervals for the mean of 100 samples taken from a normal population $N(0,1)$. As can be seen, of the 100 intervals, only 5 do not contain the true mean value $\mu = 0$.

When estimating a parameter using a confidence interval, the confidence level is usually set at high levels (the most common are 0.90, 0.95, or 0.99), to have high confidence that the parameter is within the interval. On the other hand, it is also desirable that the interval width be small to precisely delimit the value of the population parameter (this interval width is known as the *imprecision* of the estimate). However, from a sample, the higher the desired confidence level, the greater the interval width and the greater the imprecision of the estimate, and if it is required that the estimate be more precise (less imprecision), the corresponding confidence level will be smaller. Therefore, a compromise solution must be reached between the confidence level and the precision of the estimate. However, if with the available sample it is not possible to obtain an interval of sufficiently small width (small imprecision) with an acceptable confidence level, a larger sample must be used. By increasing the sample size, intervals of smaller width are obtained without decreasing the confidence level, or higher confidence levels are obtained while maintaining the width.

Confidence Intervals for the Mean

Based on conclusions drawn from the Central Limit Theorem, it is obtained that, as long as the samples are large (as a usual criterion, it is considered that they are large when the sample size, n, is greater than or equal to 30), and regardless of the original distribution of the starting variable X, with mean μ and standard deviation σ, the variable

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

follows a standard normal distribution, $N(0, 1)$.

If the standard deviation σ of the starting variable is unknown, the sample standard deviation is used as an estimate:

$$\hat{S} = \sqrt{\frac{\sum \left(X_i - \overline{X}\right)^2}{n-1}}$$

and with it, the new variable

$$T = \frac{\overline{X} - \mu}{\hat{S}/\sqrt{n}}$$

follows a Student's t distribution with $n - 1$ degrees of freedom, $T(n - 1)$.

For small samples ($n < 30$), the above results can also be applied, as long as the starting random variable X follows a normal distribution.

Based on the above and considering the three classification factors mentioned: whether the population from which the sample is taken follows a normal distribution or not, whether the variance of that population is known or unknown, and whether the sample is large ($n \geq 30$) or not, the following expressions corresponding to the different confidence intervals can be deduced.

Confidence Interval for the Mean of a Normal Population with Known Variance in Samples of Any Size

$$\left(\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\ \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

Figure 6.2 shows an explanatory diagram of the construction of this interval.

Confidence Interval for the Mean of a Normal Population with Unknown Variance in Samples of Any Size

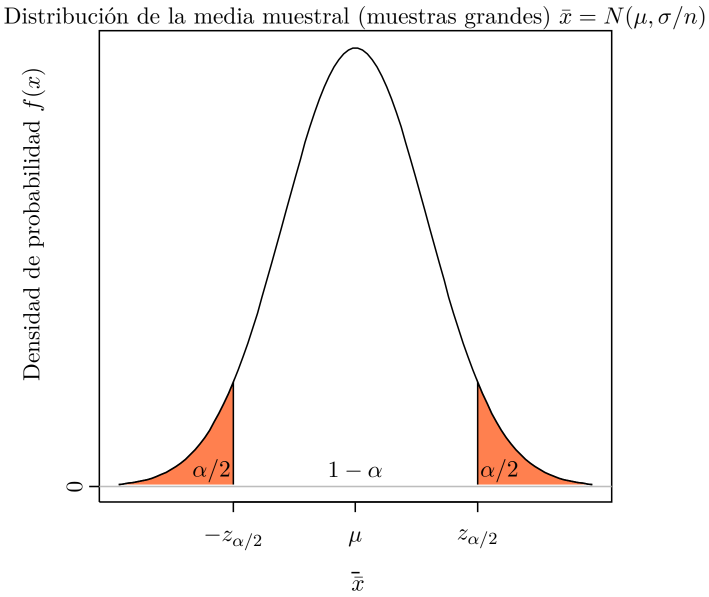
$$\left(\overline{x} - t_{\alpha/2}^{n-1} \cdot \frac{\hat{s}}{\sqrt{n}},\ \overline{x} + t_{\alpha/2}^{n-1} \cdot \frac{\hat{s}}{\sqrt{n}}\right)$$

If the samples are large ($n \geq 30$), the above interval can be approximated by:

$$\left(\overline{x} - z_{\alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}},\ \overline{x} + z_{\alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}}\right)$$

Confidence Interval for the Mean of a Non-Normal Population with Known Variance and Large Samples ($n \geq 30$)

$$\left(\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\ \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$



$$P\left(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Figure 6.2: Calculation of the confidence interval for the mean of a normal population with known variance, based on the distribution of the sample mean $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$ for samples of any size.

Confidence Interval for the Mean of a Non-Normal Population with Unknown Variance and Large Samples ($n \geq 30$)

$$\left(\overline{x} - t_{\alpha/2}^{n-1} \cdot \frac{\hat{s}}{\sqrt{n}},\ \overline{x} + t_{\alpha/2}^{n-1} \cdot \frac{\hat{s}}{\sqrt{n}}\right)$$

Since these are large samples, the above interval can be approximated by:

$$\left(\overline{x} - z_{\alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}},\ \overline{x} + z_{\alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}}\right)$$

If the starting population is not normal, and the samples are small, the Central Limit Theorem cannot be applied and confidence intervals for the mean are not obtained.

For any of the above intervals:

- n is the sample size.
- $\bar{x}$ is the sample mean.
- σ is the population standard deviation.
- $\hat{s}$ is the sample standard deviation: $\hat{s}^2 = \dfrac{\sum (x_i - \overline{x})^2}{n-1}$.
- $z_{\alpha/2}$ is the value that leaves a probability $\alpha/2$ to its right in a standard normal distribution.
- $t_{\alpha/2}^{n-1}$ is the value that leaves a probability $\alpha/2$ to its right in a Student's t distribution with $n-1$ degrees of freedom.

Confidence Intervals for the Population Proportion p

For large samples ($n \geq 30$) and values of p (probability of "success") close to 0.5, the Binomial distribution can be approximated by a Normal distribution with mean np and standard deviation $\sqrt{np(1-p)}$. In practice, for this approximation to be valid, the criterion is that both np and $n(1-p)$ must be greater than 5. This also allows us to construct confidence intervals for proportions by treating them as means of dichotomous variables in which the presence or absence of the characteristic under study ("success" or "failure") is expressed by a 1 or a 0, respectively.

Thus, in large samples and with not excessively skewed binomial distributions (both np and $n(1-p)$ must be greater than 5), if we denote $\widehat{p}$ as the proportion of individuals presenting the studied attribute in the specific sample, then the confidence interval for the proportion with a significance level α is given by:

$$\left(\widehat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\widehat{p} \cdot (1-\widehat{p})}{n}} \ , \ \widehat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\widehat{p} \cdot (1-\widehat{p})}{n}}\right)$$

where:

- n is the sample size.
- $\widehat{p}$ is the proportion of individuals presenting the studied attribute in the specific sample.
- $z_{\alpha/2}$ is the value that leaves a probability $\alpha/2$ to its right in a standard normal distribution.

In small samples or those from a strongly skewed Binomial distribution ($np \leq 5$ or $n(1-p) \leq 5$), the Central Limit Theorem cannot be applied and the construction of confidence intervals must be based on the Binomial distribution.

6.2 Solved Exercises

i. The concentration of active ingredient in a sample of 10 containers taken from a batch of a drug is analyzed, obtaining the following results in mg/mm^3:

$$17.6 - 19.2 - 21.3 - 15.1 - 17.6 - 18.9 - 16.2 - 18.3 - 19.0 - 16.4$$

A. Create the variable concentration, and enter the sample data.

B. Calculate the confidence interval for the mean concentration of the batch with a 95% confidence level (significance level $\alpha = 0.05$).

> Select the menu *Analyze→Descriptive Statistics→Explore*.
>
> In the dialog box that appears, select the variable concentration in the *Dependent List* field and click the *Statistics* button.
>
> In the dialog box that appears, enable the *Descriptives* option, enter the desired confidence level in the *Confidence Interval for Mean* field, and click the *Continue* and *OK* buttons.

C. Calculate the confidence intervals for the mean with confidence levels of 90% and 99% (significance levels $\alpha = 0.1$ and $\alpha = 0.01$).

> Repeat the same steps as in the previous section, changing the confidence level for each interval.

D. If we define the precision of the interval as the inverse of its width, how does increasing the significance level affect the precision of the confidence interval? What could be the explanation?

E. If, for the drug to be effective, it must have a minimum concentration of 16 mg/mm^3 of active ingredient, can the batch be accepted as good? Justify your answer.

ii. A dairy product plant receives milk daily from two farms X and Y. To analyze the quality of the milk, the fat content of the milk from both farms is monitored during a season, with the following results:

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 the variables fat and farm, and enter the sample data.

B. Calculate the confidence interval with 95% confidence for the average fat content of the milk without considering whether it comes from one farm or another.

> Select the menu *Analyze→Descriptive Statistics→Explore*.
>
> In the dialog box that appears, select the variable fat in the field *Dependent List* and click the button *Statistics*.
>
> In the dialog box that appears, enable the option *Descriptives*, enter the

> desired confidence level in the box *Confidence Interval for Mean*, and click the button *Continue* and *OK*.

C. Calculate the confidence intervals with 95% confidence for the average fat content of the milk by dividing the data according to the farm of origin of the milk.

> Select the menu *Analyze→Descriptive Statistics→Explore*.
>
> In the dialog box that appears, select the variable fat in the field *Dependent List*, select the variable farm in the field *Factor List* and click the button *Statistics*.
>
> In the dialog box that appears, enable the option *Descriptives*, enter the desired confidence level in the box Confidence Interval for Mean, and click the button *Continue* and *OK*.

D. In view of the intervals obtained in the previous point, can it be concluded that there are significant differences in the average fat content according to the origin of the milk? Justify the answer.

iii. In a survey conducted at a faculty, on whether students regularly (at least once a week) use the library, the following results were obtained, where 1 indicates a positive response and 0 indicates a negative response:

Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Response	0	1	0	0	0	1	0	1	1	1	1	0	1	0	1	0	0	0

Student	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
Response	1	1	1	0	0	1	0	0	1	1	0	0	1	0	1	0

A. Create the variable response and enter the sample data.

B. Calculate the confidence interval with $\alpha = 0.01$ for the mean of the response variable.

> Select the menu *Analyze→Descriptive Statistics→Explore*.
>
> In the dialog box that appears, select the variable response in the field *Dependent List* and click the button *Statistics*.
>
> In the dialog box that appears, enable the option *Descriptives*, enter the desired confidence level in the box *Confidence Interval for Mean*, and click the button *Continue* and *OK*.

C. What is the interpretation of this interval in terms of the proportion of students who regularly use the library?

iv. The Ministry of Health is interested in developing a confidence interval for the proportion of people over 65 years old with respiratory problems who have been vaccinated in a certain city. To do this, after asking 200 patients over 65 years old with respiratory problems in the hospitals of that city, 154 responded affirmatively.

A. Create the variable response whose only two values will be 0 for negative responses, and 1 for positive responses, and a second variable that we can call frequency whose values are the absolute frequencies of each of the responses (46 for response 0 and 154 for response 1).

B. Weight the values of the variable response by the weights introduced in the variable frequency.

> Select the menu *Data→Weight Cases*.
>
> In the dialog box that appears, activate the option *Weight cases by*, select the variable frequency and click the button *OK*.

C. Calculate the 95% confidence interval for the proportion of vaccinated patients.

> Select the menu *Analyze→Descriptive Statistics→Explore*.
>
> In the dialog box that appears, select the variable response in the field *Dependent List* and click the button *Statistics*.
>
> In the dialog box that appears, enable the option *Descriptives*, enter the desired confidence level in the box *Confidence Interval for Mean*, and click the button *Continue* and *OK*.

D. If one of the Ministry's objectives was to achieve a proportion of at least 70% of vaccinated individuals in this group, can it be concluded that the objectives have been met? Justify the answer.

6.3 Proposed Exercises

i. To determine the average cholesterol level in the blood of a population, analyses were performed on a sample of 8 people, obtaining the following results:

$$196 - 212 - 188 - 206 - 203 - 210 - 201 - 198$$

Find the confidence intervals for the mean cholesterol level with significance levels of 0.1, 0.05, and 0.01.

ii. Use the file Hypertensive Key Data to estimate the initial mean diastolic pressure in all individuals using a 95% confidence interval. Also calculate the 95% confidence intervals corresponding to the groups treated with placebo, ACE inhibitors, and Ca Antagonist + Diuretic. Are there significant differences in the initial mean diastolic pressure of these groups?

iii. To treat a certain neurological syndrome, two techniques A and B are used. In a study, a sample of 60 patients with this syndrome was taken, and technique A was applied to 25 of them and technique B to the remaining 35. Of the patients treated with technique A, 18 were cured, while of those treated with technique B, 21 were cured. Calculate a 95% confidence interval for the proportion of cures with each technique.

iv. In the next local elections in a city, three parties are running: A, B, and C. To make an estimate of the proportion of votes each will receive, a survey was conducted with 300 respondents, of which 60 plan to vote for A, 80 for B, 90 for C, 15 blank votes, and 55 abstentions. Calculate a confidence interval for the proportion of votes, out of the total census, for each of the parties running.

7 — Confidence Intervals for Population Comparison

7.1 Theoretical Foundations

7.1.1 Statistical Inference and Parameter Estimation

The objective of a statistical study is twofold: to describe the chosen sample from a population in which some characteristic is to be studied, and to make inferences, that is, to draw conclusions about the population from which the sample was taken.

The methodology that leads to conclusions about the population, based on the information contained in the sample, constitutes *Statistical Inference*.

Since the sample contains less information than the population, the conclusions about the population will be approximate. Therefore, one of the objectives of statistical inference is to determine the probability that a conclusion obtained from the analysis of a sample is true, and for this, it relies on probability theory.

When the value of one of the population parameters is desired, the procedure to use is *Parameter Estimation*, which is divided into *Point Estimation*, when a single value is given as an estimate of the considered population parameter, and *Interval Estimation*, when it is of interest to know not only an approximate value of the parameter but also the precision of the estimate. In this latter case, the result is an interval within which the true value of the population parameter will be, with a certain confidence. This interval is called a *confidence interval*. Unlike point estimation, where a single estimator is used, in interval estimation, two estimators are used, one for each end of the interval.

7.1.2 Confidence Intervals

Given two sample statistics L_1 and L_2, the interval $I = (L_1,\ L_2)$ is said to be a *Confidence Interval* for a population parameter θ, with *confidence level* $1-\alpha$ (or *significance level* α), if the probability that the statistics determining the interval limits take values such that θ is between them is equal to $1-\alpha$, that is,

$$P\left(L_1 < \theta < L_2\right) = 1 - \alpha$$

The ends of the interval are random variables whose values depend on the considered sample. That is, the lower and upper ends of the interval would be $L_1\left(X_1, ..., X_n\right)$ and $L_2\left(X_1, ..., X_n\right)$ respectively, although we will usually write L_1 and L_2 to simplify the notation. We will denote by l_1 and l_2 the values taken by these variables for a given sample $\left(x_1, ..., x_n\right)$.

When the definition states that the probability that the parameter θ is in the interval $(L_1,\ L_2)$ is $1-\alpha$, it means that in $100\,(1-\alpha)\ \%$ of the possible samples, the value of θ would be in the corresponding intervals $(l_1,\ l_2)$.

Once a sample is taken, and from it, the corresponding interval $(l_1,\ l_2)$ is determined, it would not make sense to talk about the probability that the parameter θ is in the interval $(l_1,\ l_2)$, since l_1 and l_2 are numbers, the parameter θ, which is also a number, although unknown, will either be in that interval or not, and therefore we talk about confidence instead of probability.

Thus, when we talk about a confidence interval for the parameter θ with a confidence level $1-\alpha$, we understand that before taking a sample, there is a probability $1-\alpha$ that the interval constructed from it contains the value of the parameter θ.

When estimating a parameter using a confidence interval, the confidence level is usually set at high levels (the most common are 0.90, 0.95, or 0.99), to have high confidence that the parameter is within the interval. On the other hand, it is also desirable that the interval width is small to precisely delimit the value of the population parameter (this interval width is known as the *imprecision* of the estimate). However, from a sample, the higher the desired confidence level, the wider the interval and the greater the imprecision of the estimate, and if it is imposed that the estimate be more precise (less imprecision), the corresponding confidence level will be smaller. Therefore, a compromise solution must be reached between the confidence level and the precision of the estimate. However, if with the available sample it is not possible to obtain an interval of sufficiently small width (small imprecision) with an acceptable confidence level, a larger sample must be used. By increasing the sample size, intervals of smaller width are obtained without decreasing the confidence level, or higher confidence levels are maintained while keeping the width.

Confidence intervals for the difference of means

Similarly to what happened with confidence intervals for the mean of a variable, based on conclusions drawn from the Central Limit Theorem, it can be shown that, in large samples ($n_1 \geq 30$ and $n_2 \geq 30$), from populations of two variables X_1 and X_2, with not necessarily Normal distributions, with means μ_1 and μ_2 and standard deviations σ_1 and σ_2 respectively, the variable

$$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

follows a standard Normal distribution, $N(0,\ 1)$.

Similarly, if the variances of the variables are unknown, using their corresponding sample variances $\hat{S}_1^2$ and $\hat{S}_2^2$ as estimators, where

$$\hat{S}_1^2 = \frac{\sum\left(X_{1,i} - \overline{X}_1\right)^2}{n_1 - 1} \quad \text{and} \quad \hat{S}_2^2 = \frac{\sum\left(X_{2,i} - \overline{X}_2\right)^2}{n_2 - 1}$$

then the variable

$$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\hat{S}_1^2}{n_1} + \dfrac{\hat{S}_2^2}{n_2}}}$$

follows a Student's t distribution, where the number of degrees of freedom depends on whether the variances, although unknown, can be considered equal or not.

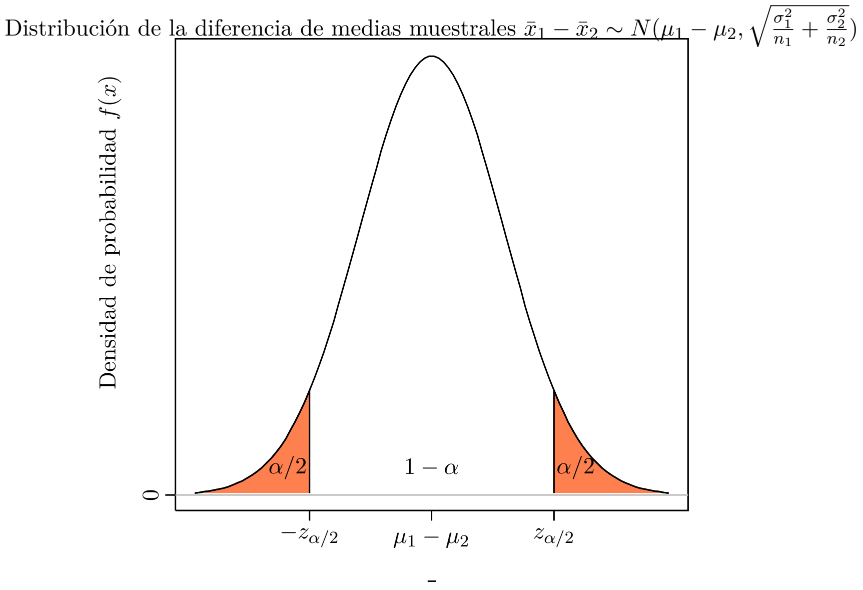
For small samples ($n_1 < 30$ or $n_2 < 30$), the above distributions are also applicable as long as the starting variables follow Normal distributions.

From all this and considering the three classification factors mentioned: whether the starting populations from which the samples are obtained follow Normal distributions or not, whether the variances of these populations are known or unknown, and whether the samples are large or not, we obtain the following expressions corresponding to the different confidence intervals.

Confidence interval for the difference of two means in normal populations, with known population variances, regardless of sample size

$$\left(\overline{x}_1 - \overline{x}_2 - z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\ ,\ \overline{x}_1 - \overline{x}_2 + z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Figure 7.1 shows an explanatory diagram of the construction of this interval.



$$P\left(\mu_1 - \mu_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \bar{x}_1 - \bar{x}_2 \le \mu_1 - \mu_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1 - \alpha$$

Figure 7.1: Calculation of the confidence interval for the difference of means in normal populations with known variances based on the distribution of the difference of sample means $\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$.

Confidence interval for the difference of two means in normal populations, with unknown population variances, regardless of sample size

If the variances, although unknown, can be considered equal, the interval is:

$$\left(\overline{x}_1 - \overline{x}_2 - t_{\alpha/2}^{n_1+n_2-2} \cdot \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\ ,\ \overline{x}_1 - \overline{x}_2 + t_{\alpha/2}^{n_1+n_2-2} \cdot \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

where s_p^2 is a weighted sample variance:

$$s_p^2 = \frac{(n_1 - 1) \cdot \hat{s}_1^2 + (n_2 - 1) \cdot \hat{s}_2^2}{n_1 + n_2 - 2}$$

If the variances, although unknown, cannot be considered equal, the interval is:

$$\left(\overline{x}_1 - \overline{x}_2 - t_{\alpha/2}^{\nu} \cdot \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}\ ,\ \overline{x}_1 - \overline{x}_2 + t_{\alpha/2}^{\nu} \cdot \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}\right)$$

where ν is the integer obtained by rounding down the value of the expression:

$$\frac{\left(\frac{\hat{s}_1^2}{n_1}+\frac{\hat{s}_2^2}{n_2}\right)^2}{\frac{\left(\frac{\hat{s}_1^2}{n_1}\right)^2}{n_1+1}+\frac{\left(\frac{\hat{s}_2^2}{n_2}\right)^2}{n_2+1}}-2$$

If the sample sizes are large ($n_1 \geq 30$ and $n_2 \geq 30$), the $t_{\alpha/2}^{\nu}$ and $t_{\alpha/2}^{n_1+n_2-2}$ can be replaced by $z_{\alpha/2}$.

Confidence interval for the difference of two means in non-normal populations, and large samples ($n_1 \geq 30$ and $n_2 \geq 30$)

In this case, as with the sample mean, the intervals for the difference of means are the same as their corresponding ones in normal populations, and again, it is necessary to distinguish whether the variances are known or unknown (equal or different), which translates into their corresponding formulas being the same as those given in the previous paragraphs. However, since these are large samples, the approximation of $t_{\alpha/2}^{\nu}$ and $t_{\alpha/2}^{n_1+n_2-2}$ by $z_{\alpha/2}$ is also valid, and usually only the distinction between known and unknown variances is made.

For known variances:

$$\left(\overline{x}_1-\overline{x}_2-z_{\alpha/2}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\ ,\ \overline{x}_1-\overline{x}_2+z_{\alpha/2}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$$

And for unknown variances:

$$\left(\overline{x}_1-\overline{x}_2-z_{\alpha/2}\cdot\sqrt{\frac{\hat{s}_1^2}{n_1}+\frac{\hat{s}_2^2}{n_2}}\ ,\ \overline{x}_1-\overline{x}_2+z_{\alpha/2}\cdot\sqrt{\frac{\hat{s}_1^2}{n_1}+\frac{\hat{s}_2^2}{n_2}}\right)$$

If the starting populations are not normal and the samples are small, the Central Limit Theorem cannot be applied, and confidence intervals for the difference of means are not obtained.

For any of the above intervals:

- n_1 and n_2 are the sample sizes.
- $\overline{x}_1$ and $\overline{x}_2$ are the sample means.
- σ_1 and σ_2 are the population standard deviations.
- $\hat{s}_1$ and $\hat{s}_2$ are the sample standard deviations: $\hat{s}_1^2 = \dfrac{\sum\left(x_{1,i}-\overline{x}_1\right)^2}{n_1-1}$, and similarly $\hat{s}_2^2$.
- $z_{\alpha/2}$ is the value that leaves a probability $\alpha/2$ to its right in a standard Normal distribution.
- $t_{\alpha/2}^{n_1+n_2-1}$ is the value that leaves a probability $\alpha/2$ to its right in a Student's t distribution with n_1+n_2-1 degrees of freedom.
- $t_{\alpha/2}^{\nu}$ is the value that leaves a probability $\alpha/2$ to its right in a Student's t distribution with ν degrees of freedom.

Confidence intervals for the mean of the difference in paired data

In many cases, it is necessary to study a characteristic in a population at two different times, to study how it evolves over time, or to analyze the impact of some event that occurred between those times.

In these cases, a random sample of the population is taken, and in each individual, the characteristic under study is observed at the two mentioned times. Thus, there are two sets of data that are not independent, as the data are paired for each individual. Therefore, the procedures seen previously cannot be applied, as they are based on the independence of the samples.

The problem is solved by taking the difference between both observations for each individual. Thus, constructing the confidence interval for the difference of means reduces to calculating the confidence interval for the mean of the difference variable. Additionally, if each set of observations follows a Normal distribution, their difference will also follow a Normal distribution.

Confidence intervals for the difference of two population proportions p_1 and p_2

For large samples ($n_1 \geq 30$ and $n_2 \geq 30$) and values of p_1 and p_2 (probability of ”success”) close to 0.5, the corresponding Binomial distributions can be approximated by Normal distributions with respective means n_1p_1 and n_2p_2, and respective standard deviations $\sqrt{n_1p_1(1-p_1)}$ and $\sqrt{n_2p_2(1-p_2)}$. In practice, for this approximation to be valid, the criterion is that both n_1p_1 and n_2p_2 as well as $n_1(1-p_1)$ and $n_2(1-p_2)$ must be greater than 5. This allows us to construct confidence intervals for the difference of proportions by treating them as means of dichotomous variables where the presence or absence of the characteristic under study (”success” or ”failure”) is expressed by a 1 or a 0 respectively.

Thus, in large samples and with not excessively skewed Binomial distributions (both n_1p_1 and n_2p_2 as well as $n_1(1-p_1)$ and $n_2(1-p_2)$ must be greater than 5), if we denote $\widehat{p}_1$ and $\widehat{p}_2$ as the proportion of individuals presenting the studied attribute in the first and second samples respectively, then the confidence interval for the difference of proportions with a significance level α is given by:

$$\left(\widehat{p}_1 - \widehat{p}_2 - z_{\alpha/2} \cdot \sqrt{\frac{\widehat{p}_1 \cdot (1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2 \cdot (1-\widehat{p}_2)}{n_2}}\ ,\ \widehat{p}_1 - \widehat{p}_2 + z_{\alpha/2} \cdot \sqrt{\frac{\widehat{p}_1 \cdot (1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2 \cdot (1-\widehat{p}_2)}{n_2}} \right)$$

where:

- n_1 and n_2 are the respective sample sizes.
- $\widehat{p}_1$ and $\widehat{p}_2$ are the proportions of individuals presenting the studied attributes in their respective samples.
- $z_{\alpha/2}$ is the value that leaves a probability $\alpha/2$ to its right in a standard Normal distribution.

In small samples or from highly skewed Binomial distributions ($n_1p_1 \leq 5$, $n_2p_2 \leq 5$, $n_1(1-p_1) \leq 5$ or $n_2(1-p_2) \leq 5$), the Central Limit Theorem cannot be applied, and the construction of confidence intervals must be based on the Binomial distribution.

Confidence interval for the ratio of two variances σ_1^2 and σ_2^2 of normal populations

As we have seen in the section on confidence intervals for the difference of two means in normal populations with unknown variances, they depend on whether the variances, although unknown, can be considered equal or not. To address this question, prior to calculating the interval for the difference of means, an interval for the ratio of variances of both populations is constructed. For this, we consider that if we start with two variables X_1 and X_2 that follow normal distributions with variances σ_1^2 and σ_2^2 respectively, and take samples of sizes n_1 and n_2 from the respective populations, the variable

$$F = \frac{\frac{\hat{S}_1^2}{\sigma_1^2}}{\frac{\hat{S}_2^2}{\sigma_2^2}}$$

follows an F distribution of Fisher with $n_1 - 1$ degrees of freedom in the numerator and $n_2 - 1$ degrees of freedom in the denominator.

From this, it follows that the confidence interval with significance level α for $\frac{\sigma_2^2}{\sigma_1^2}$ is

$$\left(\frac{\hat{s}_2^2}{\hat{s}_1^2} \cdot F_{1-\alpha/2}^{(n_1-1,n_2-1)}, \frac{\hat{s}_2^2}{\hat{s}_1^2} \cdot F_{\alpha/2}^{(n_1-1,n_2-1)}\right)$$

If the number 1 (the ratio of variances equals one) is within the obtained confidence interval, there will not be sufficient statistical evidence, with a significance level α, to reject that the variances are equal.

7.2 Solved Exercises

i. To see if an advertising campaign for a drug has influenced its sales, a sample of 8 pharmacies was taken and the number of units of the drug sold during a month was measured, before and after the campaign, obtaining the following results:

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

A. Create the variables before and after and enter the sample data.

B. Obtain a statistical summary showing the mean and standard deviation of both variables. In view of the results: are the means different? Has the campaign increased the sales level? Do you think the results are statistically significant?

> Select the menu *Analyze→Descriptive Statistics→Frequencies*.
>
> In the dialog box that appears, select the two variables in the field *Variables* and click the button *Statistics*.
>
> In the dialog box that appears, activate the desired statistics and click the button *Continue* and *OK*.

C. Obtain the confidence intervals for the mean difference between the two variables with significance levels of 0.05 and 0.01.

> Select the menu *Analyze→Compare Means→Paired-Samples T Test*.
>
> In the dialog box that appears, select both variables in the field *Paired Variables* and click the button *Options*.
>
> In the dialog box that appears, enter the desired confidence level in the field *Confidence Interval Percentage* and click the button *Continue* and *OK*.

D. Create the variable difference, which is obtained as before-after.

> Select the menu *Transform→Compute Variable*.
>
> In the dialog box that appears, enter the name of the new variable difference in the field *Target Variable*, in the field *Numeric Expression* enter before-after and click the button *OK*.

E. Calculate the confidence interval for the mean of the variable difference with a significance level of 0.05 and compare the obtained interval with the previous section.

> Select the menu *Analyze→Compare Means→One-Sample T Test*.
>
> In the dialog box that appears, select the variable difference in the field *Test Variables* and click the button *Options*.
>
> In the dialog box that appears, enter the desired confidence level in the field *Confidence Interval Percentage* and click the button *Continue* and *OK*.

F. Are there enough evidence to state with 95% confidence that the advertising campaign has increased sales? What if we change the last two data points of the variable after to 190 instead of 182 and 165 instead of 145? Observe what has happened to the interval for the mean difference and provide an explanation.

ii. A dairy product plant receives milk daily from two farms X and Y. To analyze the quality of the milk, during a season, the fat content of the milk from both farms is monitored, with the following results:

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 the variables fat and farm, and enter the sample data.

B. Calculate the confidence interval with 95% confidence for the difference in the average fat content of the milk from both farms.

> Select the menu *Analyze→Compare Means→Independent-Samples T Test*.
>
> In the dialog box that appears, select the variable fat in the field *Test Variables*, select the variable farm in the field *Grouping Variable* and click the button *Define Groups*.
>
> In the dialog box that appears, enter in the field *Group 1* the value of the variable farm corresponding to farm X and in the field *Group 2* the value corresponding to farm Y, and click the button *Continue*.
>
> In the initial dialog box, click the button *Options*.
>
> In the dialog box that appears, enter the desired confidence level in the field *Confidence Interval Percentage* and click the button *Continue* and *OK*.

C. In view of the intervals obtained in the previous point, can it be concluded that there are significant differences in the average fat content according to the origin of the milk?

D. Of the two confidence intervals obtained, one assumes the equality of variances and the other does not. In this specific example, would it be methodologically correct to consider the one that assumes equality of variances as appropriate?

> Among the results appears the Levene's test for equality of variances. It is a hypothesis test whose final result is a p-value, called Sig. by the program, which we will learn to interpret in subsequent practices. For now, it is enough to note that when the p-value is greater than the chosen significance level, we cannot reject the hypothesis of equality of variances.

iii. A university professor has had two class groups throughout the year: one in the morning and one in the afternoon. In the morning group, out of a total of 80 students, 55 passed; and in the afternoon group, out of a total of 90 students, 32 passed.

A. Create the variables: group, whose values will be 0 (morning) and 1 (afternoon); passed, whose values will be 1 (passed) and 0 (failed); and frequency, whose values will be the number of passes and fails in each group.

B. Weight the values of the variable passed by the weights of the variable frequency.

> Select the menu *Data→Weight Cases*.
>
> In the resulting dialog box, activate the option *Weight cases by*, select the variable `frequency` in the field *Frequency Variable* and click the button *OK*.

C. Calculate the interval for the difference in proportions of students who passed in each group.

> Select the menu *Analyze→Compare Means→Independent-Samples T Test*.
>
> Select the variable `passed` in the field *Test Variables*, the variable `group` in the field *Grouping Variable* and click the button *Define Groups*.
>
> In the dialog box that appears, enter in the field *Group 1* the value of the variable `group` corresponding to the morning group and in the field *Group 2* the value corresponding to the afternoon group, and click the button *Continue*.
>
> In the initial dialog box, click the button *Options*.
>
> In the dialog box that appears, enter the desired confidence level in the field *Confidence Interval Percentage* and click the button *Continue* and *OK*.

D. Assuming that other factors (syllabus, exam complexity, prior knowledge level, previous academic record, etc.) have not influenced the pass or fail in the subject, can it be concluded that the schedule factor has been decisive in the proportion of failures? Justify the answer.

7.3 Proposed Exercises

i. A study has been conducted to investigate the effect of physical exercise on blood cholesterol levels. Eleven people participated in the study, and their cholesterol levels were measured before and after completing an exercise program. The results obtained were as follows:

Previous Level	182	232	191	200	148	249	276	213	241	280	262
Post Level	198	210	194	220	138	220	219	161	210	213	226

A. Find the 90% confidence interval for the difference in the average cholesterol level before and after the exercise program.

B. In view of this interval, can it be concluded that physical exercise decreases cholesterol levels with 90% confidence?

ii. Two chemists, A and B, respectively perform 14 and 16 determinations of the radioactivity of a sample of material. Their results in Curios were:

A		B	
263.36	254.68	286.53	254.54
248.64	276.32	284.55	286.30
243.64	256.42	272.52	282.90
272.68	261.10	283.85	253.75
287.33	268.41	252.01	245.26
287.26	282.65	275.08	266.08
250.97	284.27	267.53	252.05
		253.82	269.81

A. Calculate the confidence interval for the difference in the means of the activity detected by each chemist with 95% confidence.

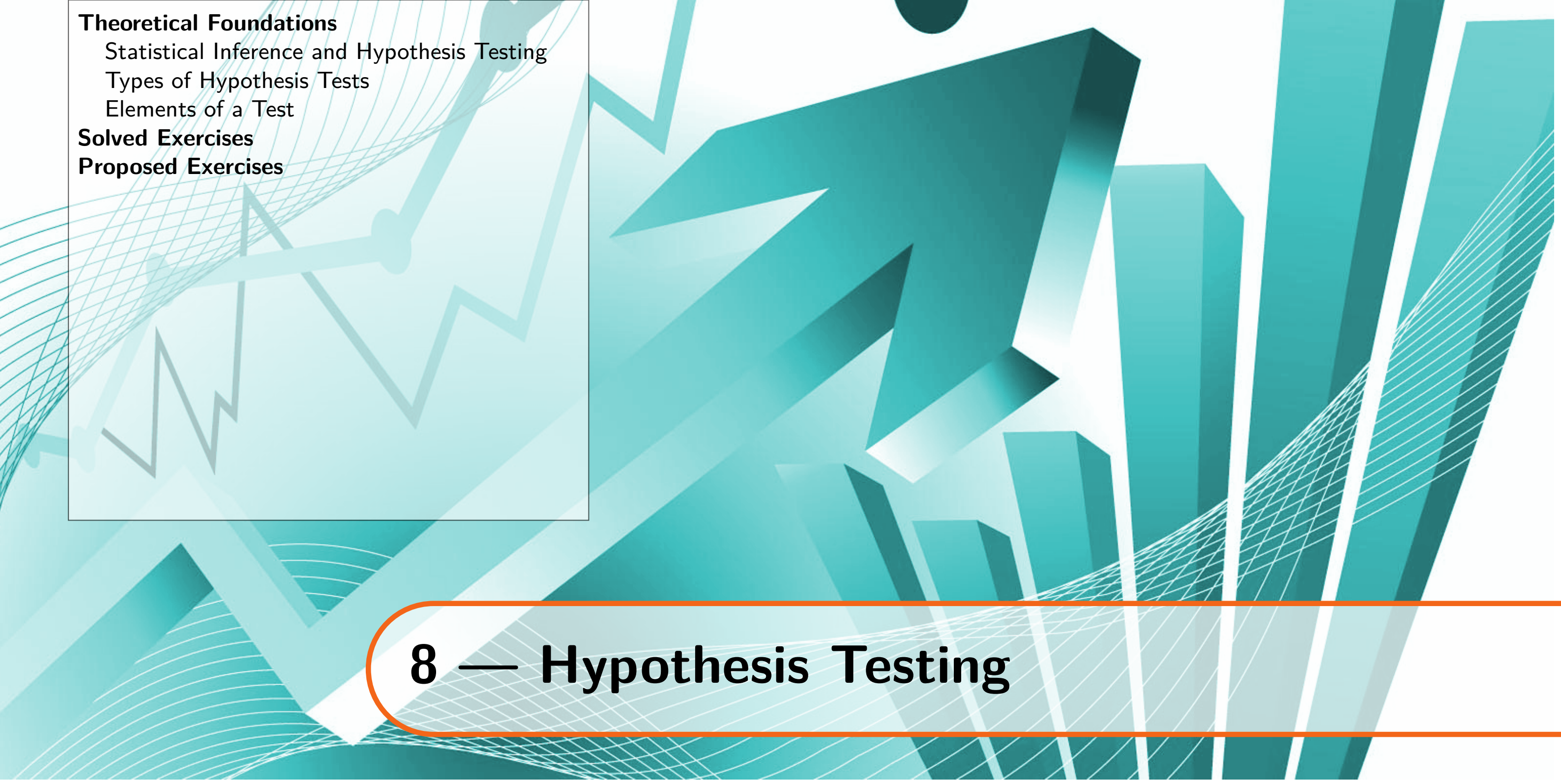
   B. Can it be said that there are significant differences in the mean activity detected by each chemist?

iii. Use the file Hypertensive Key Data to calculate the 95% confidence interval for the comparison of the initial and final mean systolic pressure in each of the treatments. Which treatment has been more effective in reducing pressure?

iv. In a survey conducted by the Junta de Comunidades de Castilla la Mancha, in the two hospitals of a city, hospitalized patients are asked upon discharge if they consider the treatment received to have been satisfactory. In the first hospital, 200 patients were asked and 140 responded positively, while in the second, 300 patients were asked and 160 responded positively.

   A. Calculate the confidence interval for the difference in proportions of patients satisfied with the treatment received.

   B. Is there significant evidence that the treatment received in one hospital is better than in the other?



8 — Hypothesis Testing

8.1 Theoretical Foundations

8.1.1 Statistical Inference and Hypothesis Testing

Any statement or conjecture that partially or completely determines the distribution of a population is made through a *Statistical Hypothesis*.

In general, it is never known with absolute certainty whether a hypothesis is true or false, as we would need to measure all individuals in the population to know for sure. Decisions are made on a probabilistic basis, and the procedures that lead to the acceptance or rejection of the hypothesis form the part of Statistical Inference known as *Hypothesis Testing*.

A hypothesis is tested by comparing its predictions with the reality obtained from the samples: if they match within the probabilistically admissible margin of error, we will maintain the hypothesis; otherwise, we will reject it and seek new hypotheses capable of explaining the observed data.

8.1.2 Types of Hypothesis Tests

Hypothesis tests are classified as:

- *Parametric Tests.* These are further divided into two types depending on whether:
  - A specific value or interval for the parameters of the distribution of a random variable is tested. For example: we can test the hypothesis that the mean blood cholesterol level in a population is 180 mg/dl.
  - The parameters of the distributions of two or more variables are compared. For example: we can test the hypothesis that the mean blood cholesterol level is lower in people who consume below a certain amount of fats in their diet.
- *Non-Parametric Tests.* These test the hypotheses imposed as a starting point in parametric tests, known as *Structural Hypotheses.* Among them, the data distribution model and the independence of the data. For example: in many parametric tests, it is required as a starting hypothesis that the sample data come from a normal population, but this would be the first test to address, since if the data do not come from a normal population, the conclusions obtained from the derived parametric tests can be completely erroneous.

8.1.3 Elements of a Test

Null Hypothesis and Alternative Hypothesis

The first step in performing a hypothesis test is the formulation of the Null Hypothesis and its corresponding Alternative Hypothesis.

We will call the *Null Hypothesis* the hypothesis being tested. It is usually represented as H_0 and represents the hypothesis that we will maintain unless the data observed in the sample indicate its falsity in probabilistic terms.

The rejection of the null hypothesis implies the implicit acceptance of the *Alternative Hypothesis*, usually represented as H_1. For each H_0, we have two different H_1 depending on whether the test is *Two-Tailed*, if we do not know the direction in which H_0 may be false, or *One-Tailed*, if we know in which direction H_0 may be false. The One-Tailed test is classified as *Right-Tailed* if H_1 only includes values greater than the parameter for which we are testing compared to those in H_0, and *Left-Tailed* if H_1 only includes smaller values.

In the following table, both H_0 and H_1 are formulated in parametric tests, for any parameter, which we will call θ, of a population, and for the comparison of two parameters, θ_1 and θ_2, of two populations.

	H_0	H_1
Two-Tailed in one population	$\theta = \theta_0$	$\theta \neq \theta_0$
One-Tailed in one population	$\theta = \theta_0$	$\theta > \theta_0$ (Right-Tailed) or $\theta < \theta_0$ (Left-Tailed)
Two-Tailed in two populations	$\theta_1 = \theta_2$	$\theta_1 \neq \theta_2$
One-Tailed in two populations	$\theta_1 = \theta_2$	$\theta_1 > \theta_2$ or $\theta_1 < \theta_2$

■ **Example 8.1** Suppose that, based on previous data, we know that the mean blood cholesterol level in a certain population is 180 mg/dl, and we assume that the application of a certain therapy may have influenced (either to increase or decrease) this mean. To formulate H_0, we must consider that the null hypothesis is always conservative, meaning we will not change our model unless there is strong probabilistic evidence that it is no longer valid. According to this, the null hypothesis will be that the mean has not changed:

$$H_0 : \mu = 180.$$

Once the null hypothesis is set, to formulate the alternative hypothesis, we must consider that it is a two-tailed test, as we do not know, a priori, the direction of the mean variation (whether it will be greater or less than 180 mg/dl). Therefore, the alternative hypothesis is that the mean is different from 180 mg/dl:

$$H_1 : \mu \neq 180.$$

On the other hand, if we assume that the application of the therapy has served to decrease the cholesterol level, we are dealing with a one-tailed test in which the null hypothesis remains that the mean has not changed, and the alternative is that it has decreased:

$$H_0 : \mu = 180,$$
$$H_1 : \mu < 180.$$

■

Normally, the researcher's goal is to reject the null hypothesis to prove the certainty of the alternative hypothesis, and this will only be done when there is sufficiently significant evidence of the falsity of H_0. If the data observed in the sample do not provide this evidence, then the null hypothesis is maintained, and in this sense, it is said to be the conservative hypothesis. However, it is important to clarify that accepting the null hypothesis does not mean it is true, but that we do not have enough information or statistical evidence to reject it.

Errors in a test. Significance level and power

As we have already mentioned, the acceptance or rejection of H_0 is always done in probabilistic terms, based on the information obtained from the sample. This means that we will never have absolute certainty about the truth or falsity of a hypothesis, so it is possible that we may make mistakes when accepting or rejecting it.

The errors that can be made in a hypothesis test are of two types:

- **Type I error**: occurs when we reject H_0 when it is true.
- **Type II error**: occurs when we accept H_0 when it is false.

The probability of making a Type I error is known as the *Significance Level* of the test and is denoted by

$$\alpha = P(\text{Reject } H_0 | H_0 \text{ is true}).$$

And the probability of making a Type II error is denoted by

$$\beta = P(\text{Accept } H_0 | H_0 \text{ is false}).$$

Thus, when performing a hypothesis test, the four situations that can occur are summarized in Table 8.1.

	Reality	
Decision	H_0 true	H_0 false
Accept H_0	Correct decision (Probability $1-\alpha$)	Type II Error (Probability β)
Reject H_0	Type I Error (Probability α)	Correct decision (Probability $1-\beta$)

Table 8.1: Types of errors in a hypothesis test.

Since the interesting aspect of a test is to reject the null hypothesis, what is most important to control is the risk of making a mistake if it is rejected, that is, the Type I error. Therefore, α is usually set at low levels, as the smaller it is, the more confidence we will have in rejecting the null hypothesis. The most common levels at which α is set are 0.1, 0.05, and 0.01.

Once the Type I error is controlled, it is also interesting to control the Type II error. However, the value of β is calculated assuming that the null hypothesis is false, that is, $\theta \neq \theta_0$ (or $\theta_1 \neq \theta_2$ in the case of two populations), but this encompasses infinite possibilities, so to calculate it, we have no choice but to fix H_1 by giving a single value to the parameter. In this case, the *Power of the Test* is defined as the probability of rejecting H_0 when the fixed alternative hypothesis is true, and it is equal to $1-\beta$. It is evident that for a given significance level, a test will be better the more power it has.

Since the power depends on the value of the parameter fixed in the alternative hypothesis, a function for the power can be defined as

$$\text{Power}(x) = P(\text{Reject } H_0 | \theta = x),$$

which indicates the probability of rejecting H_0 for each value of the parameter θ. This function is known as the *power curve.*

On the other hand, β also depends on α because as α decreases, it becomes increasingly difficult to reject H_0, and therefore, the probability of accepting the null hypothesis when it is false increases. Consequently, and as we will see later, the only way to decrease β and gain power, once α is fixed, is by increasing the sample size.

Test statistic and acceptance and rejection regions

The decision between accepting or rejecting H_0 in the test is made based on a sampling statistic related to the parameter or characteristic we want to test, and whose distribution must be known assuming H_0 is true and once the sample size is fixed. This statistic is called the *Test Statistic*.

For each sample, the test statistic takes a specific value called the *statistic estimate*. Based on this estimate, we will decide whether to accept or reject the null hypothesis. If the estimate differs too much from the value proposed by H_0 for the parameter, then we will reject H_0, while if it is not too different, we will accept it.

The magnitude of the difference we are willing to tolerate between the estimate and the parameter value to maintain the null hypothesis depends on the probability of Type I error we are willing to assume. If α is large, small differences may be sufficient to reject H_0, while if α is very small, we will only reject H_0 when the difference between the estimator and the parameter value according to H_0 is very large. Thus, by setting the significance level α, the set of values that the test statistic can take is divided into two parts: the estimates that would lead to the acceptance of H_0, called the *Acceptance Region*, and the estimates that would lead to the rejection of H_0, called the *Rejection Region*.

If we call the test statistic $\hat\theta$, then, depending on whether the test is one-tailed or two-tailed, we will have the following acceptance and rejection regions:

Test	Acceptance Region	Rejection Region
$H_0: \ \theta = \theta_0$ $H_1: \ \theta \neq \theta_0$	$\{\hat\theta_{1-\alpha/2} \leq \hat\theta \leq \hat\theta_{\alpha/2}\}$	$\{\hat\theta < \hat\theta_{1-\alpha/2}\} \cup \{\hat\theta > \hat\theta_{\alpha/2}\}$
$H_0: \ \theta = \theta_0$ $H_1: \ \theta < \theta_0$	$\{\hat\theta \geq \hat\theta_{1-\alpha}\}$	$\{\hat\theta < \hat\theta_{1-\alpha}\}$
$H_0: \ \theta = \theta_0$ $H_1: \ \theta > \theta_0$	$\{\hat\theta \leq \hat\theta_{\alpha}\}$	$\{\hat\theta > \hat\theta_{\alpha}\}$

where $\hat\theta_{1-\alpha}$ and $\hat\theta_\alpha$ are values such that $P(\hat\theta < \hat\theta_{1-\alpha} | \theta = \theta_0) = \alpha$ and $P(\hat\theta > \hat\theta_\alpha | \theta = \theta_0) = \alpha$, as shown in Figures 8.1 and 8.2.

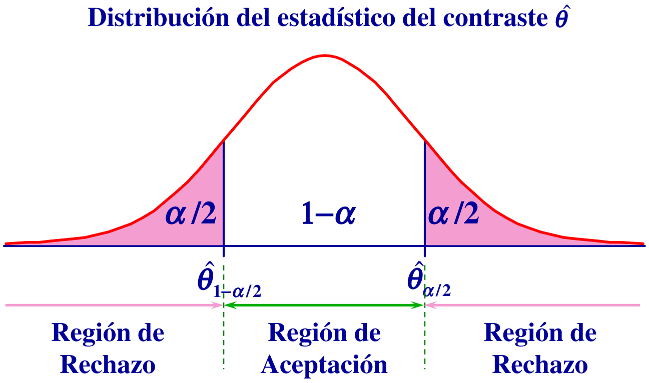


Figure 8.1: Acceptance and rejection regions in a two-tailed test.

In summary, once we have the test statistic and have set the significance level α, the acceptance and rejection regions are defined, and all that remains is to take a sample, calculate the estimate of the test statistic from it, and accept or reject the null hypothesis, depending on whether the estimate falls in the acceptance or rejection region, respectively.

The p-value of a test

Although we already have the necessary elements to perform a hypothesis test that allows us to decide whether to accept or reject the null hypothesis, in practice, the

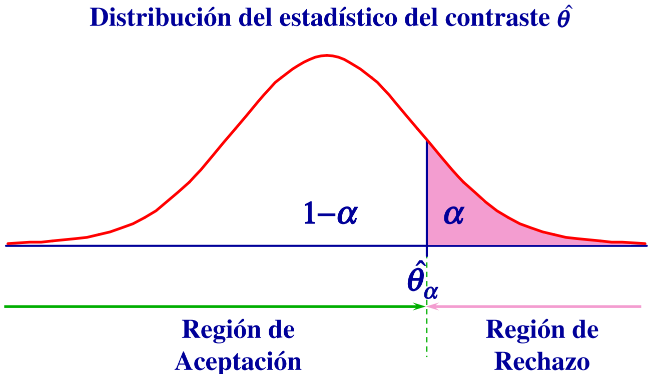


Figure 8.2: Acceptance and rejection regions in a one-tailed test.

decision is usually accompanied by the degree of confidence we have in it. For example, if we have a rejection region $\{\hat{\theta} > \hat{\theta}_\alpha\}$, whenever the estimate of the test statistic falls within this region, we will reject H_0. However, if this estimate is much greater than $\hat{\theta}_\alpha$, we will have more confidence in the rejection than if the estimate is close to the boundary between the acceptance and rejection regions $\hat{\theta}_\alpha$. For this reason, when performing a test, the probability of obtaining a discrepancy greater than or equal to the observed one between the parameter value, assuming H_0 is true, and the estimate obtained from the sample data is also calculated. This probability is known as the *p-value of the test*, and in a way, it expresses the confidence we have in making the decision in the test, since if H_0 is true and the p-value is small, it is because under the null hypothesis it is unlikely to find a discrepancy like the observed one, and therefore, we will have considerable confidence in rejecting H_0. In general, the closer p is to 1, the more confidence there is in accepting H_0, and the closer it is to 0, the more confidence there is in rejecting it.

The calculation of the p-value will depend on whether the test is two-tailed or one-tailed, and in the latter case, whether it is one-tailed to the right or one-tailed to the left. The p-value obtained for the different types of tests is shown in the following table:

Test	p-value
Two-tailed	$2P(\hat{\theta} > \hat{\theta}_0 \| H_0$ is true) if $\hat{\theta} > \hat{\theta}_0$ or $2P(\hat{\theta} < \hat{\theta}_0 \| H_0$ is true) if $\hat{\theta} < \hat{\theta}_0$
One-tailed to the right	$P(\hat{\theta} > \hat{\theta}_0 \| H_0$ is true)
One-tailed to the left	$P(\hat{\theta} < \hat{\theta}_0 \| H_0$ is true)

In Figure 8.3, it can be seen that the p-value is the area of the tail of the distribution (or tails if it is a two-tailed test) defined from the test statistic.

Once the p-value is calculated, if we have set the significance level α and the acceptance and rejection regions have been defined, the fact that the estimate falls within the rejection region is equivalent to $p < \alpha$, while if it falls within the acceptance region, then $p \geq \alpha$. This approach to tests gives us a broader view, as it provides information on which significance levels the null hypothesis can be rejected and which it cannot.

Tests and test statistics

Based on the different sampling distributions discussed in the practices on confidence intervals, the formulas for the main test statistics are presented below.

Test for the mean of a normal population with known variance

- Null Hypothesis: $H_0 : \mu = \mu_0$
- Test statistic:

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

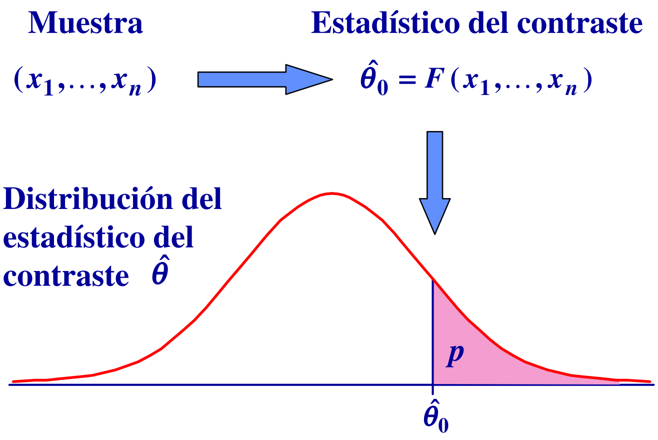

Muestra
$(x_1,\ldots,x_n)$
Estadístico del contraste
$\hat{\theta}_0 = F(x_1,\ldots,x_n)$
Distribución del estadístico del contraste $\hat{\theta}$
p
$\hat{\theta}_0$

Figure 8.3: The p-value of a one-tailed test to the right.

which follows a standard normal distribution, $N(0,1)$.

This test is also valid for the mean of a non-normal population, as long as the samples are large ($n \geq 30$), with known variance; and for the mean of the difference of paired data, as long as the difference variable follows a normal distribution with known variance, or any distribution if the sample is large.

Test for the mean of a normal population with unknown variance

- Null Hypothesis: $H_0 : \mu = \mu_0$
- Test statistic:
$$\frac{\bar{X} - \mu_0}{\hat{s}/\sqrt{n}}$$
which follows a Student's t distribution with $n-1$ degrees of freedom, $T(n-1)$.

This test is also valid for the mean of a non-normal population in large samples ($n \geq 30$), with unknown variance; and for the mean of the difference of paired data, as long as the difference variable follows a normal distribution with unknown variance, or any distribution if the sample is large.

Test for the proportion in large samples and symmetric distributions (both np and $n(1-p)$ must be greater than 5)

- Null Hypothesis: $H_0 : p = p_0$
- Test statistic:
$$\frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$
which follows a standard normal distribution, $N(0,1)$.

Test for the variance of a normal population

- Null Hypothesis: $H_0 : \sigma^2 = \sigma_0^2$
- Test statistic:
$$\frac{(n-1)\hat{s}^2}{\sigma_0^2}$$
which follows a Chi-squared distribution with $n-1$ degrees of freedom.

Test for the difference of means of normal populations with known variances

- Null Hypothesis: $H_0 : \mu_1 = \mu_2$
- Test statistic:
$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$
which follows a standard normal distribution, $N(0, 1)$.

This test is also valid for the difference of means of two non-normal populations, as long as the samples are large ($n_1 \geq 30$ and $n_2 \geq 30$), with known variances.

Test for the difference of means of normal populations with unknown variances

- Null Hypothesis: $H_0 : \mu_1 = \mu_2$
- Test statistic:
$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}}$$
which follows a Student's t distribution with ν degrees of freedom, where ν is the integer closest to the value of the expression:
$$\frac{\left(\dfrac{\hat{s}_1^2}{n_1} + \dfrac{\hat{s}_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{\hat{s}_1^2}{n_1}\right)^2}{n_1 + 1} + \dfrac{\left(\dfrac{\hat{s}_2^2}{n_2}\right)^2}{n_2 + 1}} - 2.$$

This test is also valid for the difference of means of two non-normal populations, as long as the samples are large ($n_1 \geq 30$ and $n_2 \geq 30$), with unknown variances.

Test for the difference of proportions in large samples and symmetric distributions (n_1p_1, $n_1(1 - p_1)$, n_2p_2, $n_2(1 - p_2)$ must be greater than 5)

- Null Hypothesis: $H_0 : p_1 = p_2$
- Test statistic:
$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$
which follows a standard normal distribution, $N(0, 1)$.

Test for the equality of variances of normal populations

- Null Hypothesis: $H_0 : \sigma_1^2 = \sigma_2^2$
- Test statistic:
$$\frac{\hat{S}_{1,n_1-1}^2}{\hat{S}_{2,n_2-1}^2}$$
which follows an F distribution of Fisher with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

8.2 Solved Exercises

i. The concentration of active ingredient in a sample of 10 containers taken from a batch of a drug is analyzed, obtaining the following results in mg/mm^3:

$$17.6 - 19.2 - 21.3 - 15.1 - 17.6 - 18.9 - 16.2 - 18.3 - 19.0 - 16.4$$

Tasks:

A. Create the variable concentration, and enter the sample data.

B. Perform the bilateral hypothesis test: $H_0 : \mu = 18$ and $H_1 : \mu \neq 18$ with a significance level of 0.05.

> Select the menu *Analyze→Compare Means→One-Sample T Test*.
>
> In the dialog box that appears, select the variable concentration in the field *Test Variables*, enter the value of the mean in the null hypothesis (in this case 18) in the field *Test Value*, and click the button *Options*.
>
> In the dialog box that appears, enter the confidence level of the test in the field *Confidence Interval Percentage* and click the button *Continue* and *OK*.
>
> Although the confidence level does not affect the p-value of the test, which the program calls sig. (2-tailed), it does affect the confidence interval shown along with the p-value.

C. Similarly, perform the bilateral tests: $H_0 : \mu = 19.5$ and $H_1 : \mu \neq 19.5$ with significance levels of 0.05 and 0.01. How does the decrease in the significance level affect the ease of rejecting H_0?

> Follow the same steps as in the previous section.

D. Perform the unilateral hypothesis tests: $H_0 : \mu = 17$ and $H_1 : \mu > 17$, and $H_0 : \mu = 17$ and $H_1 : \mu < 17$ with a significance level of 0.1.

> Repeat the same steps as in the previous sections but keep in mind that what the program calls Sig.(2-tailed) is the p-value of the bilateral test; therefore, for the unilateral test of greater, the p-value will be Sig.(2-tailed)/2, and for the unilateral test of lesser, the p-value will be 1-(Sig.(2-tailed)/2).

E. If the manufacturer of the batch claims to have increased the concentration of active ingredient compared to previous batches, where the mean was 17.5 mg/mm^3, with a confidence level of 95

ii. Several researchers want to know if it is possible to conclude that two populations of children differ in the average age at which they can walk on their own. The researchers obtained the following data for the age at which they started walking (expressed in months):

Sample in population A: 9.5 – 10.5 – 9.0 – 9.8 – 10.0 – 13.0 – 10.0 – 13.5 – 10.0 – 9.8

Sample in population B: 12.5 – 9.5 – 13.5 – 13.8 – 12.0 – 13.8 – 12.5 – 9.5 – 12.0 – 13.5 – 12.0 – 12.0

A. Create the variables population and age.

B. Perform a hypothesis test, with a significance level of 0.05, to answer the conclusion sought by the researchers.

> This is a bilateral test comparing means in independent populations.
>
> Select the menu *Analyze→Compare Means→Independent-Samples T Test.*
>
> In the dialog box that appears, select the variable age in the field *Test Variables*, the variable population in the field *Grouping Variable*, and click the button *Define Groups*.
>
> In the dialog box that appears, enter the value of the population variable corresponding to population A in the field *Group 1* and the value corresponding to population B in the field *Group 2*, and click the button *Continue* and *OK*.

iii. Some researchers have observed greater airway resistance in smokers than in non-smokers. To confirm this hypothesis, a study was conducted to compare the tracheobronchial retention percentage in the same people when they were still smokers and one year after quitting. The results are shown in the following table:

Retention Percentage	
When smoking	One year after quitting
60.6	47.5
12.0	13.3
56.0	33.0
75.2	55.2
12.5	21.9
29.7	27.9
57.2	54.3
62.7	13.9
28.7	8.9
66.0	46.1
25.2	29.8
40.1	36.2

A. Create the variables before and after and enter the data.

B. Formulate the appropriate hypothesis test to confirm or reject the researchers' hypothesis.

> This is a unilateral test ($H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 > \mu_2$) of equality of means in paired data.
>
> Select the menu *Analyze→Compare Means→Paired-Samples T Test.*
>
> In the dialog box that appears, select both variables and move them to the field *Paired Variables*.
>
> Since this is a unilateral test of greater, the p-value will be Sig.(2-tailed)/2.

iv. A university professor has had two class groups throughout the year: one in the morning and one in the afternoon. In the morning group, out of a total of 80 students, 55 passed; and in the afternoon group, out of a total of 90 students, 32 passed.

A. Create the variables: group, whose values will be morning and afternoon; grade, whose values will be 1 (passed) and 0 (failed); and frequency, whose different values are the number of passes and fails in each group.

B. Weight the data using the variable frequency.

> Select the menu *Data→Weight Cases.*
>
> In the resulting dialog box, activate the option *Weight cases by*, select the

> variable frequency in the field *Frequency Variable*, and click the button *OK*.

C. Perform a hypothesis test to determine whether the schedule factor has been determinant in the proportion of failures with a significance level of 0.05.

> *i*
>
> This is a bilateral test comparing means in independent populations.
>
> Select the menu *Analyze→Compare Means→Independent-Samples T Test*.
>
> Select the variable grade in the field *Test Variables*, the variable group in the field *Grouping Variable*, and click the button *Define Groups*.
>
> In the dialog box that appears, enter in the field *Group 1* the value of the variable group corresponding to the morning group and in the field *Group 2* the value corresponding to the afternoon group, and click the button *Continue*.

8.3 Proposed Exercises

i. A group of doctors is trying to prove that a moderately active exercise program can benefit patients who have previously suffered a myocardial infarction. To do this, they selected eleven individuals to participate in the study and measured their work capacity, understood as the time it takes to reach a rate of 160 beats per minute while walking at a certain speed on a treadmill, at the beginning of the study and after 25 weeks of controlled exercise. The results, expressed in minutes, were as follows:

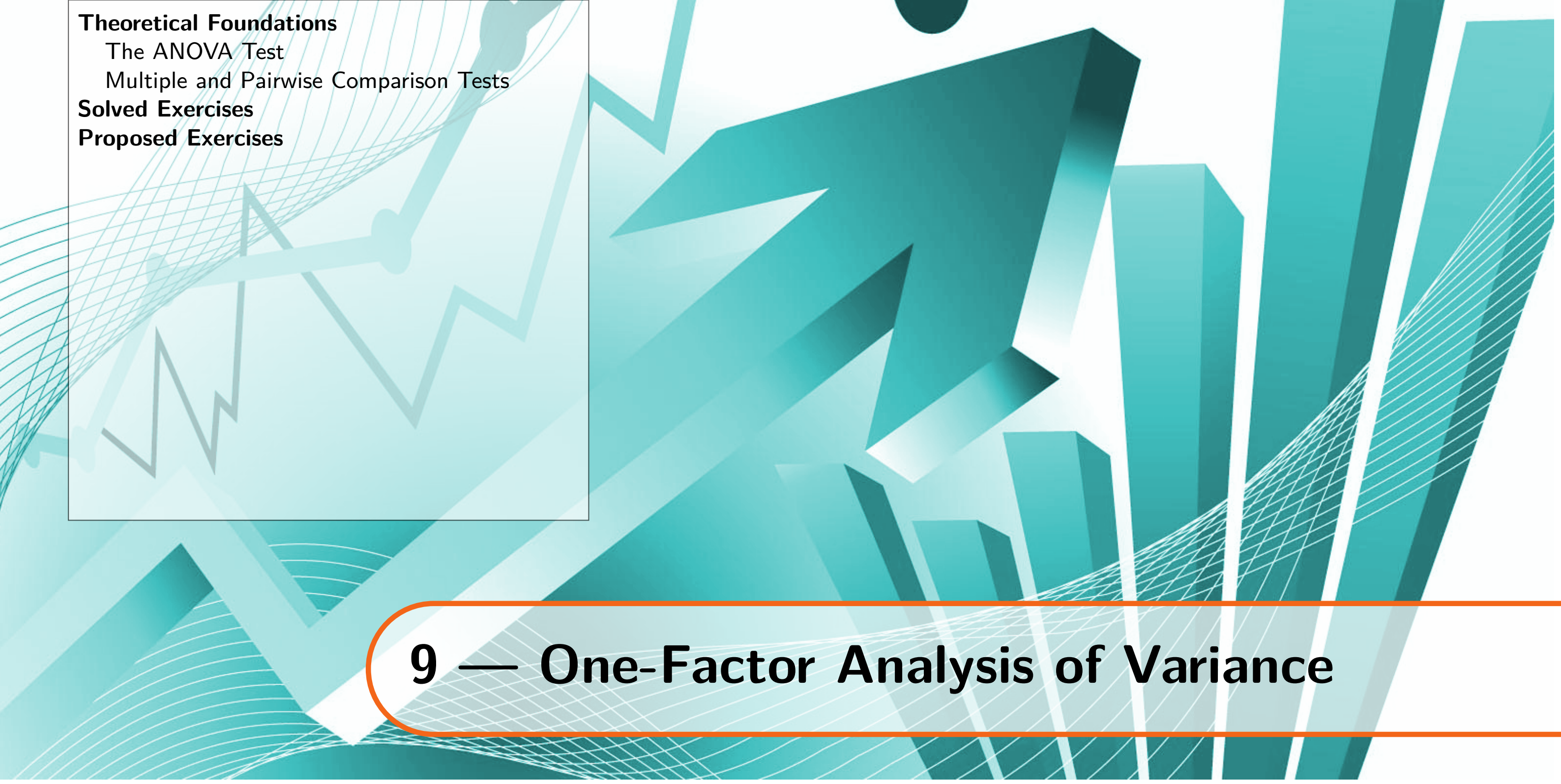
Subject	Before	After
1	7.6	14.7
2	9.9	14.1
3	8.6	11.8
4	9.5	16.1
5	8.4	14.7
6	9.2	14.1
7	6.4	13.2
8	9.9	14.9
9	8.7	12.2
10	10.3	13.4
11	8.3	14.0

Do these data support the researchers' argument?

ii. It is generally accepted that there are sex-related differences in response to heat-induced stress. To verify this, a group of 10 men and 8 women were subjected to a strenuous exercise program in a high-temperature environment (40 ºC) without the possibility of drinking. The variable of interest measured was the percentage of body weight lost. The following data were obtained:

Men		Women	
2.9	3.7	3.0	3.8
3.5	3.8	2.5	4.1
3.9	4.0	3.7	3.6
3.8	3.6	3.3	4.0
3.6	3.7		

According to the collected data, are there differences in the average percentage of body weight lost in response to physical exercise performed at high temperature between men and women?



9 — One-Factor Analysis of Variance

9.1 Theoretical Foundations

One-Factor Analysis of Variance is a statistical hypothesis testing technique used to compare the means of a quantitative variable, often called the *dependent variable* or *response*, in different groups or samples defined by a qualitative variable, called the *independent variable* or *factor*. The different categories of the factor that define the groups to be compared are known as *levels* or *treatments* of the factor.

Therefore, it is a generalization of the *t-test for comparing means of two independent samples* for experimental designs with more than two samples. It differs from a simple regression analysis, where both the dependent and independent variables were quantitative, in that in one-factor analysis of variance, the independent variable or factor is a qualitative variable.

An example of applying this technique could be comparing the average cholesterol level by blood group. In this case, the dependent variable or factor is the blood group, with four levels (A, B, O, AB), while the response variable is the cholesterol level.

To compare the means of the response variable according to the different levels of the factor, a hypothesis test is proposed in which the null hypothesis, H_0, is that the response variable has the same mean at all levels, while the alternative hypothesis, H_1, is that there are statistically significant differences between at least two of the means. This test is performed by decomposing the total variance of the response variable; hence the name of this technique: *ANOVA* (Analysis of Variance).

9.1.1 The ANOVA Test

The usual notation in ANOVA is as follows:

k is the number of levels of the factor.

n_i is the size of the random sample corresponding to the i-th level of the factor.

$n = \sum_{i=1}^{k} n_i$ is the total number of observations.

X_{ij} $(i = 1, ..., k;\ j = 1, ..., n_i)$ is a random variable indicating the response of the j-th individual at the i-th level of the factor.

x_{ij} is the specific value, in a given sample, of the variable X_{ij}.

Factor Levels			
1	2	$\cdots$	k
X_{11}	X_{21}	$\cdots$	X_{k1}
X_{12}	X_{22}	$\cdots$	X_{k2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_{1n_1}	X_{2n_2}	$\cdots$	X_{kn_k}

μ_i is the population mean of the i-th level.

$\bar{X}_i = \sum_{j=1}^{n_i} X_{ij}/n_i$ is the sample mean variable of the i-th level, and estimator of μ_i.

$\bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i$ is the specific estimate for a given sample of the sample mean variable of the i-th level.

μ is the global population mean (including all levels).

$\bar{X} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}/n$ is the sample mean variable of all responses, and estimator of μ.

$\bar{x} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}/n$ is the specific estimate for a given sample of the sample mean variable.

With this notation, we can express the response variable using a mathematical model that decomposes it into components attributable to different causes:

$$X_{ij} = \mu + (\mu_i - \mu) + (X_{ij} - \mu_i),$$

that is, the j-th response at the i-th level can be decomposed as the result of a global mean, plus the deviation from the global mean due to the fact that it receives the i-th treatment, plus a new deviation from the level mean due to random influences.

Based on this model, the null hypothesis is proposed: the means corresponding to all levels are equal; and its corresponding alternative: at least two level means are different.

$$\begin{aligned} H_0 :& \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1 :& \mu_i \neq \mu_j \text{ for some } i \neq j. \end{aligned}$$

To perform the test with this model, it is necessary to establish certain structural hypotheses (model assumptions):

Independence The k samples, corresponding to the k levels of the factor, represent independent random samples from k populations with unknown means $\mu_1 = \mu_2 = \cdots = \mu_k$.

Normality Each of the k populations is normal.

Homoscedasticity Each of the k populations has the same variance σ^2.

Considering the null hypothesis and the model assumptions, if the population means are replaced by their corresponding sample estimators in the model, we have

$$X_{ij} = \bar{X} + (\bar{X}_i - \bar{X}) + (X_{ij} - \bar{X}_i),$$

or equivalently,

$$X_{ij} - \bar{X} = (\bar{X}_i - \bar{X}) + (X_{ij} - \bar{X}_i).$$

Squaring and considering the properties of summations, we arrive at the equation known as the *sum of squares identity*:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2 = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2,$$

where:

$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X})^2$ is called the *total sum of squares* (SST), and it is the sum of squares of the deviations from the global mean; therefore, a measure of the total variability of the data.

$\sum_{j=1}^{k} n_i(\bar{X}_i-\bar{X})^2$ is called the *sum of squares of treatments or between-group sum of squares* (SSB), and it is the weighted sum of squares of the deviations of each level mean from the global mean; therefore, a measure of the variability attributed to the fact that different levels or treatments are used.

$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2$ is called the *residual sum of squares or within-group sum of squares* (SSW), and it is the sum of squares of the deviations of the observations from their respective level means; therefore, a measure of the variability in the data attributed to random fluctuations within the same level.

With this notation, the sum of squares identity is expressed as:

$$SST = SSB + SSW$$

And a final step to reach the statistic that will allow testing H_0 is the definition of *Mean Squares*, which are obtained by dividing each sum of squares by their corresponding degrees of freedom. For SST, the number of degrees of freedom is $n-1$; for SSB, it is $k-1$; and for SSW, it is $n-k$. Therefore,

$$\begin{aligned} MST &= \frac{SST}{n-1} \\ MSB &= \frac{SSB}{k-1} \\ MSW &= \frac{SSW}{n-k} \end{aligned}$$

And it can be shown that, assuming the null hypothesis and the model assumptions are true, the ratio:

$$\frac{MSB}{MSW}$$

follows an F distribution of Fisher with $k-1$ and $n-k$ degrees of freedom.

Thus, if H_0 is true, the value of the ratio for a given set of samples will be close to 0 (although always greater than 0); but if H_0 is not true, the between-group variability increases, and the statistic estimate increases. In short, we will perform a one-tailed hypothesis test with a right tail for equality of variances, and to do this, we will calculate the p-value of the obtained F estimate and accept or reject based on the fixed significance level.

ANOVA Table

All the statistics discussed in the previous section are collected in a table called the ANOVA Table, which shows the results of the estimates of these statistics for the specific samples under study. These tables are also provided as the result of any ANOVA by statistical programs, which usually add the p-value of the calculated F statistic at the end of the table, allowing us to accept or reject the null hypothesis that the means corresponding to all levels of the factor are equal.

	Sum of Squares	Degrees of Freedom	Mean Squares	F Statistic	p-value
Between Groups	SSB	$k-1$	$MSB=\frac{SSB}{k-1}$	$f=\frac{MSB}{MSW}$	$P(F>f)$
Within Groups	SSW	$n-k$	$MSW=\frac{SSW}{n-k}$		
Total	SST	$n-1$			

9.1.2 Multiple and Pairwise Comparison Tests

Once a one-way ANOVA is performed to compare the k means corresponding to the k levels or treatments of the factor, we can conclude by accepting the null hypothesis, in which case the data analysis regarding the detection of differences between levels is concluded, or by rejecting it, in which case it is natural to continue the analysis to precisely locate where the difference is, which levels have statistically different responses.

In the second case, there are several methods that allow detecting differences between the means of the different levels, known as *multiple comparison tests*. These tests are usually classified into:

Pairwise Comparison Tests Their objective is the one-by-one comparison of all possible pairs of means that can be taken when considering the different levels. The result is a table showing the differences between all possible pairs and the confidence intervals for these differences, indicating whether there are significant differences between them. It should be noted that the obtained intervals are not the same as those that would result if each pair of means were considered separately, as the rejection of H_0 in the general ANOVA test implies the acceptance of an alternative hypothesis involving several individual tests; and if we want to maintain a significance level α in the general test, in the individual tests we must use a considerably smaller α'.

Multiple Range Tests Their objective is to identify homogeneous subsets of means that do not differ from each other.

For the first type, the Bonferroni test can be used; for the second type, the Duncan test; and for both categories simultaneously, the Tukey HSD and Scheffé tests.

9.2 Solved Exercises

i. A study is conducted to compare the effectiveness of three therapeutic programs for the treatment of acne. Three methods are used:

A. Washing twice a day with a polyethylene brush and an abrasive soap, along with the daily use of 250 mg of tetracycline.

B. Application of tretinoin cream, avoiding the sun, washing twice a day with an emulsifying soap and water, and using 250 mg of tetracycline twice a day.

C. Avoiding water, washing twice a day with a lipid-free cleanser, and using tretinoin cream and benzoyl peroxide.

The study involves 35 patients. These patients were randomly divided into three subgroups of sizes 10, 12, and 13, which were assigned treatments I, II, and III, respectively. After 16 weeks, the percentage improvement in the number of lesions was recorded for each patient.

Treatment					
I		II		III	
48.6	50.8	68.0	71.9	67.5	61.4
49.4	47.1	67.0	71.5	62.5	67.4
50.1	52.5	70.1	69.9	64.2	65.4
49.8	49.0	64.5	68.9	62.5	63.2
50.6	46.7	68.0	67.8	63.9	61.2
		68.3	68.9	64.8	60.5
				62.3	

A. Create the variables treatment and improvement and enter the sample data.

> For any ANOVA, although the variable treatment is qualitative, it is advisable to enter it as quantitative, since SPSS only accepts quantitative variables as classification factors. If it is necessary to show the different levels of the factor variable as qualities, it will suffice to assign them labels.

B. Draw the scatter plot. What conclusions do you draw from the scatter plot?

> Select the menu *Graphs→Legacy Dialogs→Scatter/Dot*.
>
> In the dialog box that appears, select the option *Simple Scatter* and click the button *Define*.
>
> In the next dialog box that appears, select the variable improvement in the field *Y Axis* and the variable treatment in the field *X Axis*, and click the button *OK*.

C. Obtain the ANOVA table corresponding to the problem. Can it be concluded that the three treatments have the same average effect with a significance level of 0.05?

> Select the menu *Analyze→Compare Means→One-Way ANOVA*.
>
> In the dialog box that appears, select the variable improvement in the field *Dependent List* and the variable treatment in the field *Factor*, and click the button *OK*.

D. Obtain the ANOVA table corresponding to the problem but also show the confidence intervals of the 3 different treatments, with a significance level of 0.05, and various statistics for each of them.

> Repeat the same steps as in the previous section, clicking the button *Options* in the last dialog box and activating the option *Descriptives*.

E. If it can be concluded that the treatments have not had the same average effect, between which pairs of treatments are there statistically significant differences?

> Repeat the same steps as in the previous section, clicking the button *Post Hoc* in the last dialog box and activating the *Bonferroni* option with a significance level of 0.05.

F. If it can be concluded that the treatments have not had the same average effect, what are the homogeneous groups (groups with similar behavior in terms of improvement) of treatments that can be established?

> Repeat the same steps as in the previous section, activating the *Duncan* option.

ii. It is suspected that there are differences in the preparation for the university entrance exam among the different high schools in a city. To verify this, 8 students were randomly selected from each of the 5 schools, with the condition that they had taken the same subjects, and their scores on the university entrance exam were recorded. The results were:

Schools				
1	2	3	4	5
5.5	6.1	4.9	3.2	6.7
5.2	7.2	5.5	3.3	5.8
5.9	5.5	6.1	5.5	5.4
7.1	6.7	6.1	5.7	5.5
6.2	7.6	6.2	6.0	4.9
5.9	5.9	6.4	6.1	6.2
5.3	8.1	6.9	4.7	6.1
6.2	8.3	4.5	5.1	7.0

A. Create the variables score and school and enter the sample data.

B. Draw the scatter plot. What conclusions do you draw about the average university entrance exam score in the different schools?

> Select the menu *Graphs→Legacy Dialogs→Scatter/Dot*.
>
> In the dialog box that appears, select the option *Simple Scatter* and click the button *Define*.
>
> In the next dialog box that appears, select the variable score in the field *Y Axis* and the variable school in the field *X Axis*, and click the button *OK*.

C. Perform the ANOVA test. Can the suspicion that there are differences between the average scores of the schools be confirmed?

> Select the menu *Analyze→Compare Means→One-Way ANOVA*.
>
> In the dialog box that appears, select the variable score in the field *Dependent List* and the variable school in the field *Factor*, and click the button *OK*.

D. Which schools are the best in preparing for the university entrance exam?

Repeat the same steps as in the previous section, clicking the button *Post Hoc* in the last dialog box and activating the *Bonferroni* option to see the intervals of differences between schools, and the *Duncan* option to establish groups of homogeneous behavior.

9.3 Proposed Exercises

i. The heart rate (beats per minute) was measured in four groups of adults: normal controls (A), patients with angina (B), individuals with cardiac arrhythmias (C), and patients recovered from myocardial infarction (D). The results are as follows:

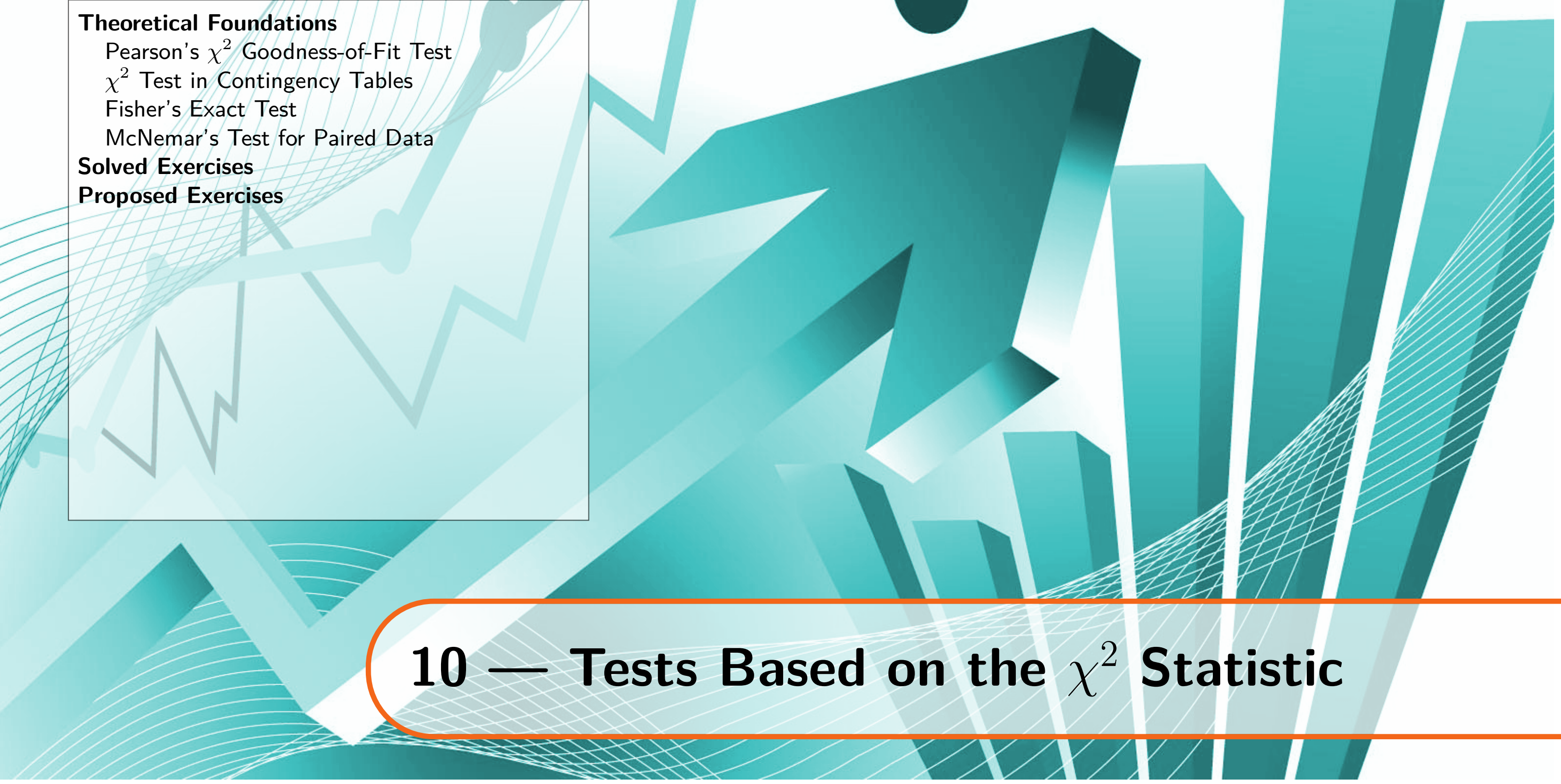
A	B	C	D
83	81	75	61
61	65	68	75
80	77	80	78
63	87	80	80
67	95	74	68
89	89	78	65
71	103	69	68
73	89	72	69
70	78	76	70
66	83	75	79
57	91	69	61

Do these data provide sufficient evidence to indicate a difference in the mean heart rate among these four types of patients? Consider $\alpha = 0.05$.

ii. The respiratory rate (breaths per minute) was measured in eight laboratory animals at three different levels of carbon monoxide exposure. The results are as follows:

Exposure Level		
Low	Moderate	High
36	43	45
33	38	39
35	41	33
39	34	39
41	28	33
41	44	26
44	30	39
45	31	29

Based on these data, is it possible to conclude that the three levels of exposure, on average, have a different effect on the respiratory rate? Take $\alpha = 0.05$.



10 — Tests Based on the χ^2 Statistic

10.1 Theoretical Foundations

There are many situations in the field of health, or in any other field, where the researcher is interested in determining possible relationships between qualitative variables. An example could be studying whether there is a relationship between complications after surgery and the patient's sex, or the hospital where the surgery is performed. In this case, all the inference techniques seen so far for quantitative variables are not applicable, and for this, we will use a hypothesis test based on the χ^2 (Chi-square) statistic.

However, although this is its most well-known aspect, the use of the test is not limited to studying the possible relationship between qualitative variables, and it is also applied to check the fit of the sample distribution of a variable, whether qualitative or quantitative, to its hypothetical theoretical distribution model.

In general, this type of test involves taking a sample and observing if there is a significant difference between the *observed frequencies* and those specified by the theoretical model being tested, also known as *expected frequencies*.

We could say that there are two main blocks of basic applications in the use of the χ^2 test:

i. **Goodness-of-fit test**. It is a significance test to determine if the data from the population, from which we have taken a sample, conform to a theoretical distribution law that we suspect is correct.

   For example: we have 400 data points that, a priori, follow a uniform probability distribution, but is it statistically certain that they fit this type of distribution?

ii. **Test for contingency tables.** These start from the two-dimensional frequency table for the different modalities of the qualitative variables. Although very often the χ^2 test applied in contingency tables is called an independence test, it is actually applied in two different experimental designs, which classify it into two different blocks:

   A. **Independence test**. Through which the researcher aims to study the relationship between two qualitative variables in a population.

      For example: we have a sample of 200 patients (the researcher only controls the total in one sample) operated on for appendicitis in 4 different hospitals, and we want to see if there is a relationship between possible postoperative infection and the hospital where the patient was operated on.

B. **Homogeneity test**. Through which the researcher aims to see if the proportion of a certain characteristic is the same in possibly different populations.

For example: we have two different samples, one of 100 HIV-positive individuals and another of 600 HIV-negative individuals (the researcher controls the total in both samples), and we want to analyze if the proportion of individuals with gastrointestinal problems is the same in both.

Finally, although the Chi-square test is very important in analyzing relationships between qualitative variables, its application can lead to errors in certain situations; especially when sample sizes are small, which leads to having very few individuals in some categories, invalidating the test's assumptions; and also when we have qualitative variables analyzed in the same individuals but at different times, that is, through paired data. For the first case, when the number of individuals in some category is very small, the Fisher's Exact Test is used, while in the second case, with paired data, the McNemar's Test is used.

10.1.1 Pearson's χ^2 Goodness-of-Fit Test

It is the oldest goodness-of-fit test and is valid for all types of distributions. To analyze a sample of a variable grouped into categories (even if it is quantitative), evaluating a prior hypothesis about the probability of each modality or category, a Chi-square goodness-of-fit hypothesis test is performed.

The test is based on counting the data and comparing the observed frequencies of each modality with the expected frequencies according to the theoretical model being tested. Thus, the statistic is calculated as:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

where O_i are the observed frequencies in the sample in modality i, and E_i are the expected frequencies for the same modality according to the theoretical model. The expected frequencies are calculated by multiplying the sample size by the probability of the corresponding modality according to the theoretical model, that is, $E_i = np_i$, where p_i is the probability of modality i.

If the population from which the sample was taken follows the theoretical distribution model, the above statistic follows a χ^2 distribution with $k-1$ degrees of freedom, where k is the number of modalities of the variable. A large value of the χ^2 statistic indicates that the distributions of the observed and expected frequencies are quite different, while a small value of the statistic indicates that there is little difference between them.

The χ^2 goodness-of-fit test is valid if all expected frequencies are greater than or equal to 1 and no more than 20% of them have expected frequencies less than 5. If this condition is not met, then the involved categories should be combined with adjacent categories to ensure that all meet the condition. If the categories correspond to categorized quantitative variables, they do not necessarily have to correspond to the same variable amplitude.

10.1.2 χ^2 Test in Contingency Tables

As we have seen, the χ^2 test in contingency tables is used to establish relationships between qualitative variables (or categorized quantitative variables), for which regression and correlation analysis cannot be performed, and both to determine independence between variables and homogeneity between populations (equal proportion of a certain characteristic). To do this, we describe the methodological process in the case of independence between variables, which in practice, and although conceptually different, is the same for homogeneity between populations.

Contingency tables are understood as those double-entry tables where the sample is classified according to a double classification criterion. For example, classifying individuals according to their sex and blood group would create a table where each cell of the table would represent the bivariate frequency of the characteristics corresponding to its row and column (for example, women with blood group A). If a random sample of size n is taken in which both variables are measured and the frequencies of the observed pairs are represented in a two-dimensional table, we have:

X/Y	y_j	
x_i	n_{ij}	n_i
	n_j	n

Where n_{ij} is the absolute frequency of the pair (x_i, y_j), n_i is the marginal frequency of the modality x_i and n_j is the marginal frequency of the modality y_j. These frequencies appear in the margins of the contingency table by summing the frequencies by rows and columns, and are therefore known as marginal frequencies.

Following a procedure similar to the previous section, the observed frequencies in the sample (real frequencies) are compared with the expected frequencies (theoretical frequencies). To do this, we calculate the probability of each cell in the table considering that if both variables are independent, the probability of each cell arises as a product of probabilities (probability of the intersection of two independent events) $p_{ij} = p_i p_j = \frac{n_i}{n}\frac{n_j}{n}$. In this way, we obtain the expected frequency as:

$$E_i = np_{ij} = n\frac{n_i}{n}\frac{n_j}{n} = \frac{n_i n_j}{n},$$

And with this, the Pearson's Chi-square statistic is calculated:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

In the case that X and Y are independent, this statistic follows a Chi-square distribution with $(f-1)(c-1)$ degrees of freedom, where f is the number of rows in the contingency table and c the number of columns. A large value of the Chi-square statistic indicates that the distributions of the observed and expected frequencies are quite different, and therefore lack of independence; while a small value of the statistic indicates that there is little difference between them, which indicates that they are independent.

This test is appropriate if the expected frequencies for each cell are at least 1 and no more than 20% of them have expected frequencies less than 5. In the case of a 2x2 table, these figures are reached only when no expected frequency is less than 5. If this is not met, among other things, a test for small samples called Fisher's exact test can be used.

10.1.3 Fisher's Exact Test

This test can be used when the necessary conditions for applying the Chi-square test are not met (more than 20% of the expected frequencies for each cell are less than 5). Although, given the large number of calculations necessary to reach the final result of the test, statistical programs only calculated it for 2x2 contingency tables, in current versions they can incorporate specific modules that extend the capacity of the base program core, and that do allow evaluating the Fisher's test in tables with more categories. For example, in PASW, if the Exact Tests module is installed, Fisher's Exact Test can be calculated in general tables with F categories in the rows and C categories in the columns.

Fisher's Exact Test is based on the exact distribution of the data and not on asymptotic approximations, and assumes that the marginals of the contingency table are fixed. The procedure for its calculation consists of evaluating the associated probability, under the assumption of independence, for all the tables that can be formed with the same marginal totals as the observed data and varying the frequencies of each cell to consider all situations in which there is a disproportion as large or larger than in the analyzed table. For the calculation of the associated probability of each table, the probability function of a hypergeometric discrete variable is used.

Although generally Fisher's Exact Test is more conservative than the Chi-square (it is more difficult to detect statistically significant differences between proportions), it has the advantage that it can be applied without any restriction on the frequencies of the cells in the contingency table.

10.1.4 McNemar's Test for Paired Data

So far we have assumed that the samples to be compared were independent, that is, two different groups in which a certain characteristic was observed. Therefore, we have made comparisons of proportions of individuals who present a certain characteristic in two different groups, but we can also consider comparing the proportion of individuals who present that characteristic in the same group of individuals but analyzed at two different times. In this latter case, it is called a comparison of proportions in paired data.

For example, if we want to see if there are differences in the improvement of symptoms of a certain disease, and for this we apply two different drugs to a group of individuals at two different times when they have contracted the same disease. In this case, it might seem appropriate to apply both the Chi-square and Fisher's exact test to determine if there are differences between the two drugs in the proportion of cured patients, but here there is a fundamental difference with the previous cases, which is that we only have one group of patients and not two. In this type of study, random variability is considerably reduced, as it is the same individual who undergoes both treatments, and the improvement in symptoms will not depend on other factors as important as, for example, age, sex, or type of diet, which may influence but may not be adequately controlled in an independent group design. By reducing random variability through paired data, small differences between proportions can become significant, even with small sample sizes, which means that this type of experimental design is more efficient in obtaining statistically significant results.

However, new designs imply new ways of handling data, and the most appropriate procedure is the one used in the McNemar's test for paired data. For its application in our example, a table with 4 cells should be constructed in which the people who have improved symptoms with both drugs, those who have improved with the first and not with the second, those who have improved with the second and not with the first, and those who have not improved with either should be counted.

Improvement 1st \ Improvement 2nd	Yes	No	Totals
Yes	a	b	$a+b$
No	c	d	$c+d$
Totals	$a+c$	$b+d$	$n=a+b+c+d$

With this, the sample proportion of patients who have experienced improvement with drug 1 is: $\widehat{p}_1 = (a+b)/n$, and similarly with drug 2: $\widehat{p}_2 = (a+c)/n$, and we can formulate the test whose null hypothesis is that there is no difference in population proportions between the two drugs: $H_0 : p_1 = p_2$, which can be performed by considering the appropriate confidence interval for the difference in proportions, or also that, under the assumption of equal proportions

$$z = \frac{b-c}{\sqrt{b+c}},$$

is a statistic that follows a standard normal distribution, and

$$\chi^2 = \frac{(b-c)^2}{b+c},$$

is a statistic that follows a Chi-square distribution with one degree of freedom. With either of them, the p-value of the test can be calculated.

10.2 Solved Exercises

i. Given two pairs of genes Aa and Bb, the offspring of the cross performed according to Mendel's laws should be composed as follows:

Phenotype	Relative Frequencies
AB	$9/16 = 0.5625$
Ab	$3/16 = 0.1875$
aB	$3/16 = 0.1875$
ab	$1/16 = 0.0625$

Choosing 300 individuals at random from a certain population, the following frequency distribution is observed:

Phenotype	Observed Frequencies
AB	165
Ab	47
aB	67
ab	21

A. Create the variables phenotype and frequency and enter the sample data.

B. Weight the data using the variable frequency.

> Select the menu *Data→Weight Cases*.
>
> In the resulting dialog box, activate the option *Weight cases by*, select the variable frequency in the *Frequency Variable* field, and click the *OK* button.

C. Check if this sample complies with Mendel's laws.

> Select the menu Analyze→Nonparametric Tests→Legacy Dialogs→Chi-Square.
>
> In the dialog box that appears, select the variable phenotype in the *Test Variable List* field, and in Expected Values mark the option *Values* and enter the proportions according to Mendel's laws following the order in which the phenotypes appear, and click the *OK* button.

D. Based on the test results, can it be accepted that Mendel's laws are followed in the individuals of this population?

ii. In a study on peptic ulcers, the blood group of 1655 ulcer patients and 10000 controls was determined, the data were:

	O	A	B	AB
Patient	911	579	124	41
Controls	4578	4219	890	313

A. Create the variables participants, blood_group, and frequency and enter the data.

B. Weight the data using the variable frequency.

> Select the menu *Data→Weight Cases*.
>
> In the resulting dialog box, activate the option *Weight cases by*, select the variable frequency in the *Frequency Variable* field, and click the *OK* button.

C. Construct the contingency table and perform the Chi-square test.

> Select the menu *Analyze→Descriptive Statistics→Crosstabs*.
>
> In the dialog box that appears, select the variable participants in the *Rows* field and the variable blood_group in the *Columns* field, and click the *Statistics* button.
>
> In the dialog box that appears, check the Chi-Square box and click the *Continue* button.
>
> In the initial dialog box, click the *Cells* button.
>
> In the dialog box that appears, check the Observed and Expected boxes, and click the *Continue* and OK buttons.

D. Based on the test results, is there any relationship between blood group and peptic ulcer? That is, can it be concluded that the proportion of patients and controls is different depending on the blood group?

iii. Mitchell et al. (1976, Annals of Human Biology), based on a sample of 478 individuals, studied the distribution of blood groups in various regions of southwestern Scotland, obtaining the results shown:

	Eskdale	Annandale	Nithsdale	
A	33	54	98	185
B	6	14	35	55
O	56	52	115	223
AB	5	5	5	15
	100	125	253	478

A. Create the variables blood_group, region, and frequency and enter the data.

B. Weight the study by the variable frequency

> Select the menu *Data→Weight Cases*.
>
> In the resulting dialog box, activate the option *Weight cases by*, select the variable frequency in the *Frequency Variable* field, and click the *OK* button.

C. Construct the contingency table and perform the Chi-square test.

> Select the menu *Analyze→Descriptive Statistics→Crosstabs*.
>
> In the dialog box that appears, select the variable blood_group in the *Rows* field and the variable region in the *Columns* field, and click the *Statistics* button.
>
> In the dialog box that appears, check the Chi-Square box and click the *Continue* button.
>
> In the initial dialog box, click the *Cells* button.
>
> In the dialog box that appears, check the Observed and Expected boxes, and click the *Continue* and OK buttons.

D. Based on the test results, are blood groups distributed in the same way in the different regions?

iv. In a study to determine if the habit of smoking is related to gender, 26 people were surveyed. Of the 9 men consulted, 2 responded that they smoked, while of the 17 women consulted, 6 smoked. Can we say that there is a relationship between the two variables?

A. Create the variables gender, smokes, and frequency and enter the data.

B. Weight the study by the variable frequency

> Select the menu *Data→Weight Cases*.
>
> In the resulting dialog box, activate the option *Weight cases by*, select the variable frequency in the *Frequency Variable* field, and click the *OK* button.

C. Construct the contingency table and perform the Chi-square test.

> Select the menu *Analyze→Descriptive Statistics→Crosstabs*.
>
> In the dialog box that appears, select the variable gender in the *Rows* field and the variable smokes in the *Columns* field, and click the *Statistics* button.
>
> In the dialog box that appears, check the Chi-Square box and click the *Continue* button.
>
> In the initial dialog box, click the *Cells* button.
>
> In the dialog box that appears, check the Observed and Expected boxes, and click the *Continue* and OK buttons.

D. Based on the test results, are smokers distributed in the same way in both genders?

> In this case, the procedure to follow is the same as for the Chi-square, but we see that now the conditions to apply this test are not met, so we will have to look at the Fisher's Exact Test, which we can apply, considering whether we are performing a bilateral or unilateral test.

v. To test the effectiveness of two different drugs against migraines, 20 people who regularly suffered from migraines were selected, and each was given the drugs at different times. They were then asked if they had experienced improvement with the drug taken. The results were as follows:

	1	2	3	4	5	6	7	8	9	10
Drug 1	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes
Drug 2	No	No	Yes	No	Yes	Yes	No	No	No	No

	11	12	13	14	15	16	17	18	19	20
Drug 1	Yes	No	Yes	No	Yes	Yes	Yes	No	Yes	Yes
Drug 2	Yes	No	Yes	No	No	Yes	No	Yes	No	No

A. Create the variables Improvement_Drug1, and Improvement_Drug2 and enter the data.

B. Construct the contingency table and perform the McNemar test.

> Select the menu *Analyze→Descriptive Statistics→Crosstabs*.
>
> In the dialog box that appears, select the variable Improvement_Drug1 in the *Rows* field and the variable Improvement_Drug2 in the *Columns* field, and click the *Statistics* button.
>
> In the dialog box that appears, check the McNemar box and click the *Continue* button.
>
> In the initial dialog box, click the *Cells* button.
>
> In the dialog box that appears, check the Observed and Expected boxes, and click the *Continue* and OK buttons.

C. Based on the test results, can we say that there are significant differences between the two drugs?

> Another way to perform this same test is by selecting the menu *Analyze →Nonparametric Tests→Legacy Dialogs→2 Related Samples*. Then move the two variables to be tested to the Test Pairs box, check the McNemar box, and click the OK button.

10.3 Proposed Exercises

i. Suppose we want to check if a die is well balanced or not. We roll it 1200 times, and we get the following results:

Number	Frequencies of occurrence
1	120
2	275
3	95
4	310
5	85
6	315

A. Based on the results, can we accept that the die is well balanced?

B. We are told that, in this die, even numbers appear with a frequency 3 times higher than odd numbers. Test this hypothesis.

ii. A study is conducted in a population of hypothetical critical patients, observing, among other things, two variables: the outcome (whether they survive SV or not NV) and the presence or absence of coma at admission. The following results are obtained:

	No coma	Coma	
SV	484	37	521
NV	118	89	207
	602	126	728

We ask ourselves: is coma at admission a risk factor for mortality?

iii. The recovery produced by two different treatments A and B is classified into three categories: very good, good, and poor. Treatment A is administered to 32 patients and B to another 28. Of the 22 very good recoveries, 10 correspond to treatment A; of the 24 good recoveries, 14 correspond to treatment A, and of the 14 poor recoveries, 8 correspond to treatment A. Are both treatments equally effective for patient recovery?

iv. To test the hypothesis that women are more successful in their studies than men, a sample of 10 boys and another of 10 girls who have been examined by a teacher who always passes 40% of the students presented for the exam is taken. Considering that only 2 boys passed, use the most appropriate hypothesis test to decide if the mentioned hypothesis is true.

v. 150 students of a course were asked if they agreed or not with the teaching methodology of two different professors who taught them in the biostatistics subject. The results are shown in the following table:

Professor 1 \ Professor 2	Favorable opinion	Unfavorable opinion
Favorable opinion	37	48
Unfavorable opinion	44	21

Can we say that there is a different opinion among the students about the two professors?