People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research

University of Constantine 03
Salah Boubnider
Faculty of Medicine
Department of Medicine

Course Handout and Practical Works

Introduction to Biomedical Data Analysis

-Biostatistics–Informatics-

Intended for First-Year Medical Students

Written by: Dr. Abdennour Boulesnane

Academic Year: 2025-2026

Preface

Dear Students,

I am pleased to present to you this course handout dedicated to the Biostatistics and Informatics module for first-year medical students. In this module, you will find a comprehensive introduction to the analysis of biomedical data and its statistical processing using Excel and SPSS.

The content of this module is highly relevant to the current scientific challenges in medicine and computational medicine. The effective use of medical data and the application of statistical and computational methods are essential for advancing the diagnosis and treatment of diseases.

That is why this handout is designed to provide you with a solid foundation in statistics and informatics, which is essential for medical research, enabling you to make informed decisions and to conduct high-quality research.

The purpose of this handout is to provide you with a clear and comprehensive learning resource to facilitate your studies. Each chapter begins with a summary of the key points to remember, followed by a detailed explanation of the concepts covered, along with numerous examples to help you fully understand the material presented. The practical exercises included will allow you to apply these concepts and acquire the skills necessary to succeed in our field.

In the following seven chapters, you will learn how to collect your data, preprocess it using various techniques, perform advanced statistical analyses such as analyzing relationships between variables, and work with probability distributions using SPSS. Each chapter is accompanied by practical examples to help you better understand the concepts and techniques discussed.

We hope that this handout will provide you with a rewarding and stimulating learning experience. We wish you all an excellent reading and great success in your studies!

Sincerely,

Dr. Abdennour Boulesnane
Last updated on : March 10, 2026

Contents

List of Figures

List of Tables

Chapter 1

Biomedical Data Analysis

1.1 Introduction

Biomedical data analysis is important because it allows researchers to identify disease patterns, develop predictive models, facilitate drug development, enable personalized medicine, and integrate data from multiple sources to provide a more comprehensive view of patient health. Through these applications, biomedical data analysis plays a crucial role in advancing our understanding of diseases and in the development of new treatments and therapies to improve patient outcomes.

1.2 Data Science

1.2.1 Definition

Data science is an interdisciplinary field that involves extracting, processing, analyzing, visualizing, and interpreting large and complex datasets. It uses a combination of statistical and computational methods to discover patterns and insights from data that can inform decision-making and provide a competitive advantage.

Data science involves a range of skills, including statistical analysis, machine learning, data visualization, and data management. It often entails working with large and diverse datasets from various sources, including social media, e-commerce transactions, sensors, medical records, and financial data.

The goal of data science is to transform data into actionable information that can guide decision-making, improve efficiency, and provide a competitive edge. Data scientists use a range of tools and techniques to achieve this goal, including data cleaning and preprocessing, exploratory data analysis, feature engineering, model selection and validation, and data visualization.

Data science has a wide range of applications across all sectors, including healthcare, finance, marketing, social media, and e-commerce. Examples of data science applications include fraud detection, recommendation systems, personalized marketing, predictive maintenance, and medical diagnosis.

1.2.2 Data Science Methodology

The methodology of data science is a systematic approach to solving data-driven problems that involves several key steps (see Figure 1.1):

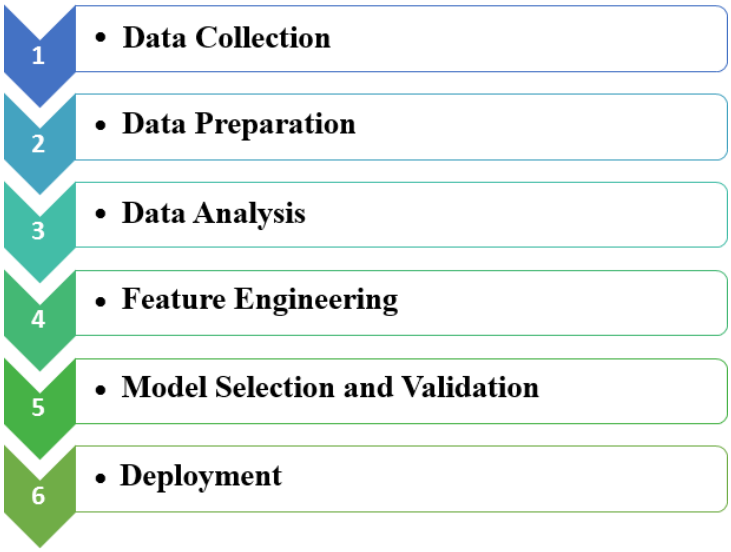


Figure 1.1: Flowchart of the Data Science Methodology

1. **Data Collection**: This involves identifying relevant data sources, gathering the data, and storing it in a structured format.
2. **Data Preparation**: This involves cleaning, transforming, and preprocessing the data to make it suitable for analysis.
3. **Data Analysis and Exploratory Data Analysis (EDA)**: EDA and data analysis are related but distinct approaches to analyzing data.
   EDA is a preliminary step in data analysis that uses various statistical and visualization techniques to explore and understand the data. The goal of EDA is to discover patterns, trends, and anomalies in the data that can inform the data analysis process. EDA involves techniques such as histograms, boxplots, scatter plots, and correlation matrices to visually explore the data and identify key features and relationships.
   Data analysis, on the other hand, is a more formal and structured process that involves applying statistical and machine learning techniques to the data. The goal of data analysis is to extract insights and make predictions from the data using a well-defined methodology. Data analysis involves techniques such as hypothesis testing, regression analysis, classification, and clustering to model and interpret the data.
   Although EDA and data analysis are distinct processes, they are often used together in the data analysis workflow.
4. **Feature Engineering**: This involves selecting and transforming relevant features or variables in the data to create new features that improve model performance.
5. **Model Selection and Validation**: This involves selecting appropriate statistical or machine learning models, training them on the data, and evalu-

ating their performance using suitable metrics and validation techniques.

6. **Deployment**: This step involves deploying the model into a production environment and integrating it into the business workflow.

The data science methodology is an iterative process that involves cycling back through these steps until a satisfactory solution is found. It requires a combination of technical skills, domain knowledge, and creativity to develop effective solutions for complex, data-driven problems.

1.3 Computational Medicine

1.3.1 Definition

Computational medicine is an interdisciplinary field that combines principles from computer science, mathematics, and engineering with medicine to develop computational tools and models for solving health-related problems. It involves the use of data-driven approaches, computer modeling, and simulation to improve medical diagnosis, treatment, and patient care.

The goal of computational medicine is to provide personalized and precise medical care through the use of data-based models and simulations. These models can help identify risk factors, predict disease progression, and optimize treatment plans. For example, machine learning algorithms can be used to analyze medical images to detect early signs of cancer or identify patterns in large datasets that can inform more effective treatment strategies.

Some applications of computational medicine include drug discovery and development, medical imaging and diagnostics, predictive modeling, disease classification, and personalized medicine. Computational medicine is a rapidly growing field with the potential to transform healthcare by providing more accurate and efficient diagnostic and treatment methods.

1.3.2 Application Areas

Computational medicine has a wide range of applications across various fields, including:

1. **Medical Imaging**: Analyzing medical images (MRI, CT, X-rays) to assist in diagnosis, treatment planning, and disease monitoring.
2. **Genomics and Personalized Medicine**: Using genomic data to identify disease-causing mutations, predict individual risk, and tailor treatment plans.
3. **Electronic Health Records (EHRs)**: Mining EHRs to identify disease patterns, predict patient outcomes, and guide personalized care.
4. **Drug Discovery and Development**: Identifying new drug targets, designing and optimizing molecules, and predicting drug toxicity.
5. **Clinical Trials**: Designing and analyzing trials to increase efficiency, monitor adverse events, and improve trial outcomes.

6. **Public Health**: Tracking epidemics, forecasting disease spread, and developing prevention and control strategies.
7. **Medical Devices and Sensors**: Optimizing devices and wearable sensors for diagnosis, monitoring, and treatment.
8. **Health Informatics**: Developing systems to improve healthcare delivery and patient outcomes.
9. **Telemedicine and Remote Monitoring**: Analyzing data from telehealth consultations and wearable devices to monitor patients remotely.
10. **Predictive Analytics for Hospital Management**: Using data to forecast patient admissions, optimize staffing, and improve hospital resource allocation.
11. **Clinical Decision Support Systems (CDSS)**: Developing tools that help clinicians make evidence-based decisions at the point of care.
12. **Epidemiology and Population Health**: Using computational models to understand disease trends in populations and plan preventive interventions.
13. **Behavioral and Lifestyle Analysis**: Integrating data from apps, wearables, and questionnaires to study lifestyle factors and their impact on health.
14. **Rare Disease Research**: Using computational methods to identify patterns in rare diseases where sample sizes are limited.
15. **Biomedical Literature Mining**: Extracting insights from the vast biomedical literature to inform research and clinical decisions.

By exploring these diverse applications, medical students can better appreciate how computational methods and data science directly impact patient care, research, and healthcare innovation.

1.4 Data Science in Computational Medicine

Data science plays a crucial role in computational medicine. The abundance of medical data generated from various sources, including electronic health records, medical imaging, genomics, and wearable devices, has created a need for data-driven approaches to analyze and interpret this information.

Data science methods, such as machine learning, statistical modeling, and data mining, can be used to extract patterns and insights from medical data that inform diagnosis, treatment, and patient care. For example, machine learning algorithms can be applied to analyze medical images to detect early signs of cancer or to identify patterns in large datasets that can be used to develop more effective treatment plans.

In computational medicine, data science can be used to develop predictive models capable of forecasting disease progression or predicting patient outcomes. Data science can also help identify risk factors and tailor treatment

plans based on patient-specific data, such as genetic information, medical history, and lifestyle factors.

Overall, data science is a critical component of computational medicine, enabling researchers and clinicians to leverage the vast amounts of medical data available to improve diagnosis, treatment, and patient care.

1.4.1 Biomedical Data

Biomedical data refers to any type of data related to human health and biology. It can include a wide range of information, such as patient medical records, clinical trial data, genomic data, imaging data, and many other types of health-related information.

Biomedical data is generally used to better understand disease mechanisms, identify potential treatments, and improve patient care. With the rise of big data analysis and artificial intelligence, the use of biomedical data has attracted growing interest for developing predictive models and personalized medicine approaches.

However, biomedical data also raises significant ethical and privacy concerns, particularly given the sensitive nature of the data and the potential for misuse. As such, strict regulations govern the collection, storage, and use of biomedical data to ensure it is handled responsibly and ethically.

1.4.2 Biomedical Data Analysis Tools

There are numerous tools available for analyzing biomedical data, including:

1. **Statistical Software**: Statistical software such as R, SAS, Python, and **SPSS** are commonly used to analyze biomedical data, including clinical trials and epidemiological studies.
2. **Data Visualization Tools**: Tools such as Tableau, MATLAB, and **Excel** can be used to create visualizations of biomedical data, including charts and diagrams, to identify patterns and trends.
3. **Machine Learning and Artificial Intelligence Tools**: Tools such as TensorFlow and Keras can be used to develop predictive models and identify patterns in large biomedical datasets.
4. **Genomic Analysis Tools**: Tools such as GATK, SAMtools, and Picard can be used to analyze genomic data, including DNA sequencing and gene expression data.
5. **Imaging Analysis Tools**: Software such as ImageJ and OsiriX can be used to analyze medical imaging data, including CT scans, MRIs, and X-rays.
6. **Network Analysis Tools**: Tools such as Cytoscape and Gephi can be used to analyze complex biological networks and identify relationships among different components.

7. **Text Mining and Natural Language Processing Tools**: Tools such as PubMed and MetaMap can be used to extract information from biomedical literature and electronic health records.

These are just a few examples of the many tools available for analyzing biomedical data. The choice of tool depends on the specific needs of the researcher and the nature of the data being analyzed.

1.4.3 Excel

Excel is a widely used spreadsheet software for data analysis and management in many fields, including biomedical research. While Excel can be useful for certain types of biomedical data analysis, it may not be the most appropriate software for all types of biomedical datasets.

One limitation of Excel is its maximum number of rows and columns, which can be a constraint when working with large datasets. Additionally, Excel lacks built-in features for advanced statistical analysis or sophisticated data visualization, which can limit its utility for certain types of biomedical data analysis.

However, Excel can be useful for basic data management tasks such as data entry, sorting, filtering, and basic calculations. It can also help generate simple tables and charts for data visualization. Furthermore, Excel can be used to track experimental data or create basic spreadsheets to calculate simple statistical measures, such as mean, median, and standard deviation.

1.4.4 Statistical Package for the Social Sciences (SPSS)

SPSS, which stands for Statistical Package for the Social Sciences, is software used for statistical analysis in the social sciences, including psychology, sociology, and related fields. It is a popular tool for analyzing research data and is widely used in both academic and commercial settings.

SPSS offers a user-friendly interface for performing statistical analyses, allowing researchers to easily enter data, run analyses, and generate tables and charts to present results. Some statistical tests that can be performed using SPSS include t-tests, ANOVA, regression analysis, factor analysis, and cluster analysis.

SPSS also provides a wide range of data management and transformation tools, such as sorting, merging, and recoding data, which are helpful for preparing data for analysis.

Moreover, SPSS can be useful for certain types of biomedical data analysis, such as survey data analysis or performing statistical tests on clinical trial data. It provides an intuitive interface for data entry and analysis and can generate tables and charts to present results clearly.

Overall, SPSS is a powerful and versatile tool for statistical analysis, especially for those working in data science. It is designed to be user-friendly, even for those with limited statistical knowledge, and can produce accurate and reliable results.

1.5 Conclusion

Biomedical data analysis and computational medicine are rapidly evolving fields that are transforming healthcare and medical research. By leveraging large and complex datasets, advanced statistical methods, machine learning, and computational models, researchers and clinicians can gain deeper insights into disease mechanisms, improve diagnostic accuracy, optimize treatment plans, and advance personalized medicine.

The integration of data science techniques into computational medicine enables predictive modeling, pattern recognition, and evidence-based decision-making, ultimately enhancing patient outcomes. Tools such as **Excel** and **SPSS** play a crucial role in these processes, providing accessible platforms for data management, statistical analysis, and visualization to support research and clinical decision-making. Therefore, mastering these tools and methodologies of biomedical data analysis, is essential for modern medical research and practice. These skills empower researchers and clinicians to extract meaningful insights from data and contribute to the advancement of healthcare.

Chapter 2

Data Collection with Excel

2.1 Introduction

Data collection is a fundamental component of biomedical research, as it helps generate the information and knowledge necessary to advance our understanding of diseases and develop new treatments and therapies. Without proper data collection, it would be difficult to progress biomedical research and improve patient outcomes.

In biomedical research, Excel is commonly used for data collection. Excel is a versatile tool that allows users to create customized spreadsheets to store and organize data in a structured manner. It also offers features such as data validation, formulas, and templates that help ensure data accuracy, streamline data analysis, and standardize the data collection process.

2.2 Starting Excel

There are different ways to start Excel:

- **Start Menu (Windows 7)**: Click the Start button ➜ All Programs ➜ Microsoft Office ➜ Microsoft Office Excel 2013.
- **Start Menu (Windows 11)**: Click the Start button ➜ Type "Excel" in the search bar ➜ Select Microsoft Excel from the search results.
- **Shortcut**: To make launching Excel easier, it is recommended to create a shortcut on the Desktop. Then, double-click this shortcut to start Excel.

2.3 Terminology

Excel, Workbook, Worksheet, Column, Row, Cell, Cell Reference, Range of Cells (see Figure 2.1).

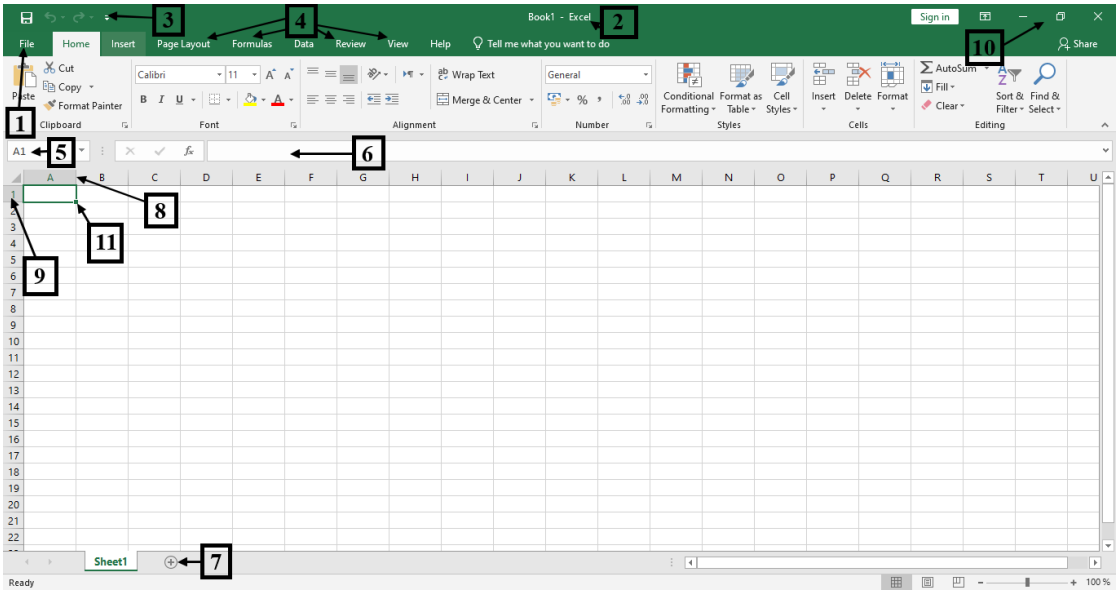


Figure 2.1: Excel Environment.

- An Excel file is called a Workbook. It can contain multiple worksheets **(7)**. By default in Excel 2013, it contains one worksheet: Sheet1. Each worksheet consists of columns **(8)** and rows **(9)**. The intersection of these columns and rows forms cells.
- In Excel, each column is referenced by one, two, or three letters (A, B, C, ..., XFD) — a total of 16,384 columns. Rows are numbered from 1 to 1,048,576. A cell reference is obtained by combining the column letter(s) with the row number (without any space). For example, the intersection of the 4th column (D) and the 6th row gives cell D6.
- By default, cells are empty, but they can contain values such as text, numbers, or formulas.

2.4 Excel Window

■ **File Tab (1)**: In Excel 2013, the Office Button was replaced by the File tab, located in the upper-left corner of the window. Clicking it opens the Backstage view, which contains two main panels. The right panel shows a list of recently used workbooks, while the left panel provides frequently used commands such as "New", "Open", "Save", "Print", etc.

■ **Title Bar (2)**: Displays the name of the current workbook followed by the application name (Excel). On the right, there are three buttons **(10)**: Minimize, Maximize/Restore, and Close.

■ **Quick Access Toolbar (3)**: Contains buttons for frequently used commands, accessible without switching tabs. By default, it shows Save, Undo, and Redo, but it can be customized to include commands like New, Open, or Quick Print.

■ **Ribbon**: Located below the title bar, the Ribbon is organized hierarchically

into multiple tabs **(4)** and contextual tabs. Fixed tabs include Home, Insert, Page Layout, Formulas, Data, Review, and View. Contextual tabs appear when an object is selected. For example, selecting a chart displays the Chart Tools contextual tabs : Design, Layout, and Format. Each tab is divided into groups containing command buttons and galleries. For example, the Home tab contains groups: Clipboard, Font, Alignment, Number, Styles, Cells, and Editing. Note: You can collapse the Ribbon by double-clicking the active tab (**[Ctrl] + [F1]**):

- To temporarily display the groups, click on a tab.
- To restore the Ribbon, double-click the tab again.

■ **Name Box (5)**: Displays the address or name of the active cell (currently selected) or selection. Only one cell can be active at a time, highlighted with a thick border.

■ **Formula Bar (6)**: Displays the content of the active cell and allows entry or editing of cell content.

■ **Fill Handle (11)**: The small green square at the lower-right corner of a cell. When hovered, the pointer becomes a black cross and can be used to copy or extend cell contents.

2.5 XLS or XLSX?

See Table 2.1.

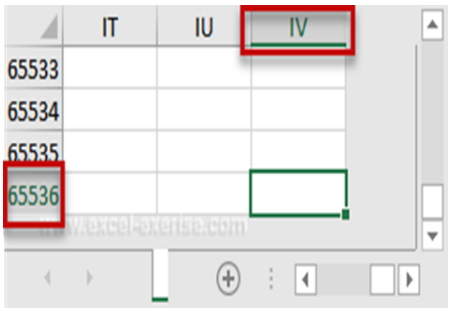
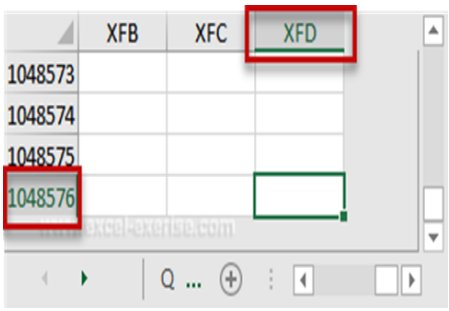
Excel 97-2003	**Excel 2007 and later (2010, 2013, 2016, 2019, 2021, 365)**
In an XLS workbook, the limits are 65,536 rows and 256 columns, which corresponds to column IV.	In an XLSX workbook, the limits are 1,048,576 rows and 16,384 columns, which corresponds to column XFD.
	

Table 2.1: XLS vs XLSX in modern Excel versions.

2.6 Workbook Management

2.6.1 Create a Workbook

To create a new workbook **[Ctrl]+[N]**: Click **File Tab** ⇨ New ⇨ Blank Workbook.

2.6.2 Open a Workbook

To open a workbook from Excel **[Ctrl]+[O]**: Click **File Tab** ⇨ Open.

2.6.3 Save a Workbook

Excel offers several ways to save a new workbook (physically creating it) on your hard drive:

1. Click **File Tab** ⇨ Save (**[Ctrl]+[S]**)
2. Click the floppy disk icon  on the Quick Access Toolbar (**[Ctrl]+[S]**)
3. Click **File Tab** ⇨ Save As (**[F12]**)

When using one of these save methods (1, 2, or 3) for the first time, the "Save As" dialog appears, where you specify:

- ✓ The file name
- ✓ The drive and folder for saving
- ✓ The file type: the desired format for saving. By default, workbooks are saved in the (.xlsx) format (Excel 2007 and later, including 2013)

2.7 Active Cell and Range of Cells

■ The **active cell** is the cell in which data will be entered. It is distinguished by a thicker border. By default, cell A1 is the active cell when a workbook is opened. The address (or name) and the content of the active cell are displayed in the Name Box and the Formula Bar, respectively.

■ Any rectangular block of cells is called a **range of cells**, or simply a **range**. To refer to a range of cells, it is common to use the reference of the top-left cell followed by a colon and the reference of the bottom-right cell.

Example:
- ● A1:B3 refers to the cells: A1, B1, A2, B2, A3, B3
- ● C1:E3 refers to the cells: C1, D1, E1, C2, D2, E2, C3, D3, E3
- ● A1:A8 refers to the cells: A1, A2, A3, A4, A5, A6, A7, A8
- ● A1:E1 refers to the cells: A1, B1, C1, D1, E1

2.8 Data Entry

Before entering data into a cell, you must first select it.

❖ Text

■ Text is automatically aligned to the left.
■ Text does not wrap automatically even if it exceeds the column width.
■ To wrap text within a cell, press **[Alt]+[Enter]** inside the cell.
■ If text begins with "+", "-", or "=", Excel will display an error message "($\#NAME?$)" because it interprets the text as a formula. To avoid this, add an apostrophe ' before the text (example: '`+Medicine`').
■ Add an apostrophe ' before a number to have it interpreted as text (example: '`2019`' in cell A2).

❖ Number

■ Numbers are automatically aligned to the right.
■ To enter a negative number, either precede it with "-" or enclose it in parentheses.
■ Writing 32e9 means $32 * 10^9$, i.e., 32 followed by nine zeros.

❖ Date

■ Dates are automatically aligned to the right.
■ To enter a date in a cell, type it as: dd/mm/yyyy or dd-mm-yyyy.
■ To enter today's date, press **[Ctrl]+[;]** (this date is fixed).
■ To have a date interpreted as text, precede it with an apostrophe ' (example: '`01/09/2025`) (see Figure 2.2).

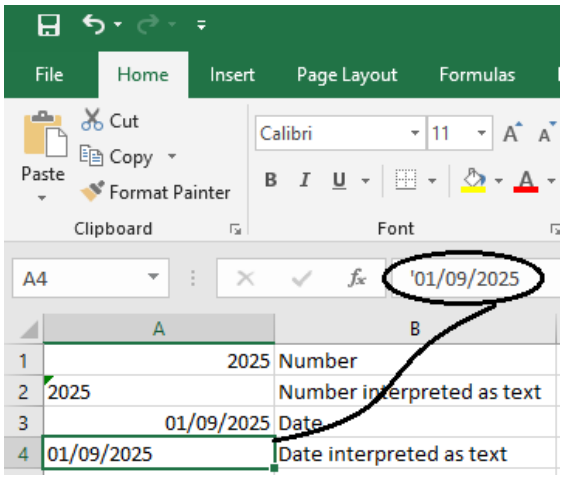


Figure 2.2: Using the apostrophe to enter text in Excel.

2.9 Data Series

➤ Enter `Patient` in cell A1 and `Date` in cell B1.
➤ Select cell A1 and drag the fill handle (small square at the bottom-right corner of the cell) down to fill several cells with consecutive entries (Figure 2.3).
➤ For dates in column B, Excel can automatically increment the values by day, month, or year depending on the pattern you establish.

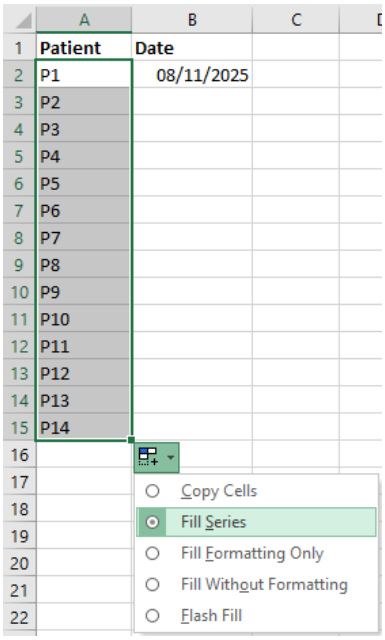


Figure 2.3: Using the fill handle in Excel to complete a data series.

2.10 Validation of Statistical Variables

Variables are used to describe individuals in a population. Each column corresponds to a variable.

Table 2.2: Quantitative and Qualitative Variables

Types of Variables			
Quantitative (Numeric)		**Qualitative (Categorical)**	
Continuous	**Discrete**	**Ordinal**	**Nominal**
Composed of numeric values that can be measured, but not counted (infinite).	Composed of numeric values that can be measured and counted (finite).	Composed of text or labels that have a logical order.	Composed of text or labels without logical order.
Example: Weight {56.06 kg, 87 kg}	Example: Number of children {0, 1, 2, 3,..., 10}	Example: Tumor size {small, medium, large}	Example: Gender {Male, Female}

■ A variable has a name: "ID", "Age", "Weight", ...
■ A variable has a value at a certain point: MED01, 22 years, 59 kg (each row represents an individual, statistical unit, ...)

Note:
Just because a variable is represented by numbers does not mean it is automatically quantitative. Numeric values may simply serve as labels for qualitative categories. For example, if we code "Male" as 1 and "Female" as 2, these numbers only identify distinct categories and do not indicate a measurable relationship between them.

2.10.1 Quantitative Variables

■ Using the "Data Validation" window, it is possible to restrict input by imposing rules such as allowing only whole numbers.
Example: For the variable `Age`: To restrict input to integers between 0 and 100, first select the relevant cells (before entering any values), for example C2:C6 (step S1). Then go to **Data** → **Data Tools** → **Data Validation** (S2 and S3). Choose **Allow: Whole number** (S4). Then in **Data: between**, type **0** in the **Minimum** box and **100** in the **Maximum** box (S5) (see Figure 2.4).

Notes:
■ For continuous quantitative variables, choose **Decimal** in the "Allow" list.
■ In the "Data" list, you have several options such as: **greater than**, **less than**, **between**, etc.

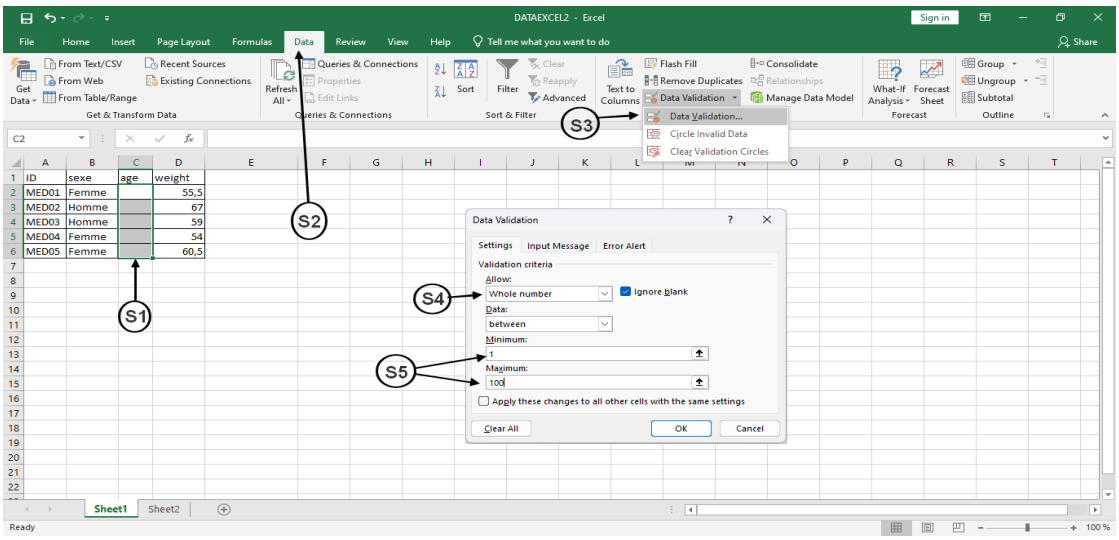


Figure 2.4: Data validation for quantitative variables

2.10.2 Qualitative Variables

■ Select the cells where you will enter `Sex` (step E1 in the following figure) (in our example B2:B6) (see Figure 2.5).

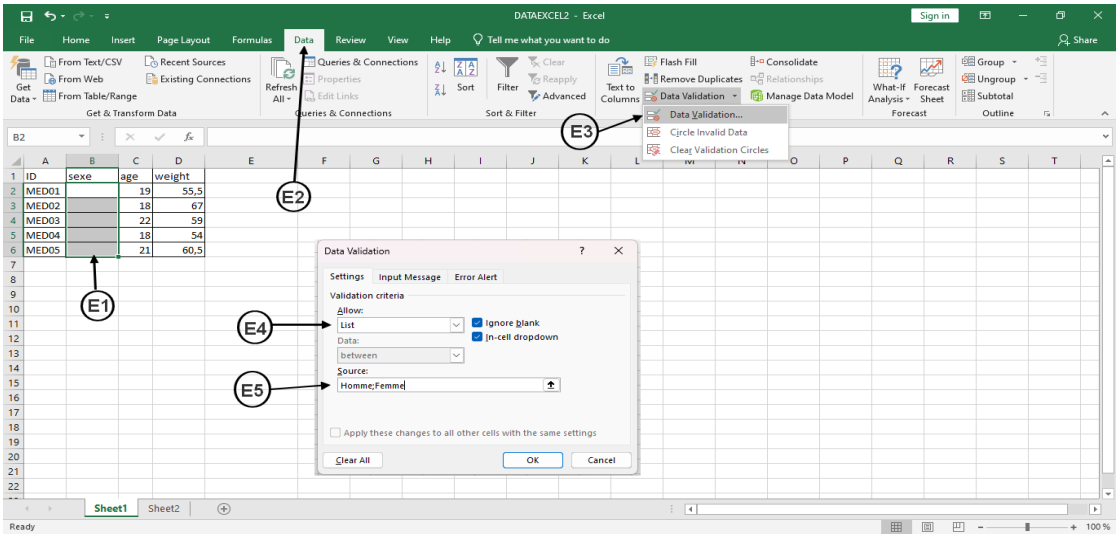


Figure 2.5: Data validation for qualitative variables

Next, choose **Allow: List** (E4) under **Data** → **Data Validation** → **Data Validation** (E2 and E3). Enter the list items (**Homme; Femme**) in the Source field (E5), separated by a semicolon **;**.

The result will be:

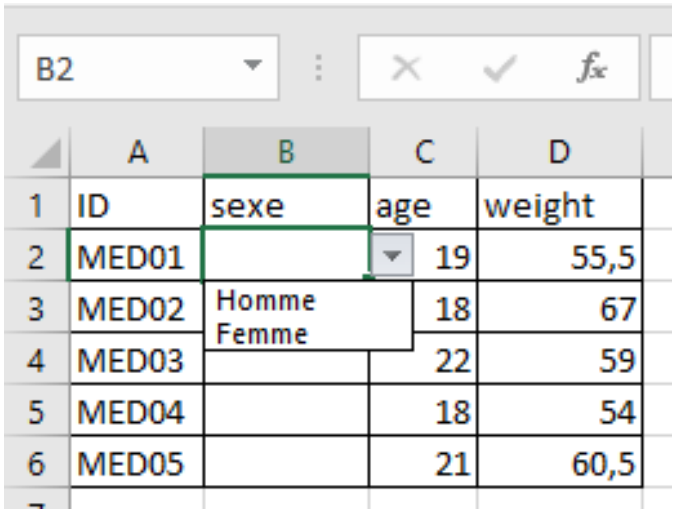


Figure 2.6: Result of data validation for qualitative variables

Note:
■ You can find the Excel data file for this chapter at: `https://aboulesnane.net/wp-content/datafiles/DATAEXCEL2.xlsx`

2.11 Conclusion

Excel is a popular tool for data collection in biomedical research. To collect data efficiently with Excel, it is important to plan your spreadsheet, use data validation to ensure accuracy, apply formulas to streamline analysis, and secure your data. Consequently, Excel can be a powerful tool to organize and analyze data in biomedical research.
In summary, mastering Excel is an essential skill for biomedical researchers and students, as it supports accurate data collection, facilitates initial data exploration, and lays the groundwork for advanced analyses and reproducible research.

Chapter 3

Organization of the Collected Data in SPSS

3.1 Introduction

SPSS (Statistical Package for the Social Sciences) is a powerful software tool widely used to organize, analyze, and visualize data in social and biomedical research. It offers a range of customizable options for data organization and management, powerful statistical tools for data analysis, and clear and transparent documentation for reproducibility.

3.2 The SPSS

The SPSS environment consists of several components, as shown in the image below (Figure 3.1):

1. The title bar: displays the name of the current file and the application.
2. The menu bar: This bar provides access to various commands grouped according to their function. SPSS has a number of menu options located at the top of the screen (like any other computer program). Open SPSS and select each menu option one by one.
    - **The 'File' menu (shortcut Alt + F)**: Essentially, this menu allows you to open existing files, create new ones, and print or save whatever you are working on. The Recently Used Data and Recently Used Files lists are useful because they allow you to quickly access the files you have recently opened or worked on.
    - **The 'Edit' menu (shortcut Alt + E)**: This menu should be familiar if you have used word processors before. Undo and Redo can help correct mistakes you make. Cut, Copy, and Paste allow you to move blocks of numbers from one part of the spreadsheet to another. Find... and Go to Case... allow you to locate a particular data score or participant, which is very convenient when dealing with a large amount of data.

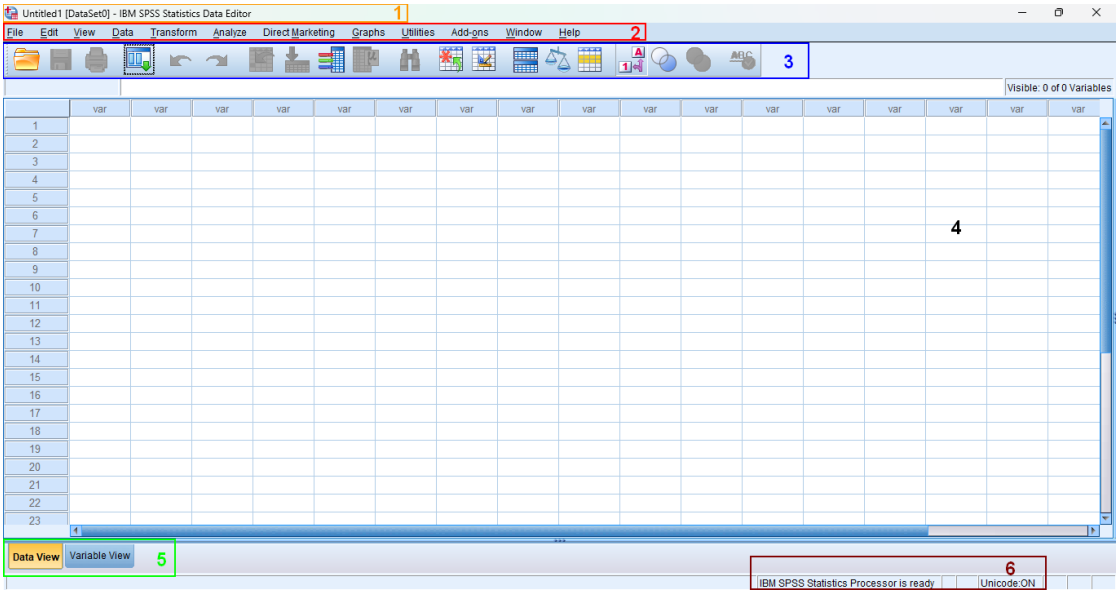


Figure 3.1: SPSS Environment.

- **The 'View' menu (shortcut Alt + V)**: The View menu deals with visual aspects of the spreadsheet, particularly: which toolbars are displayed, which fonts are used, whether grid lines are visible on the spreadsheet, whether value labels are displayed for your variables..., etc.
- **The 'Data' menu (shortcut Alt + D)**: This menu allows you to organize your data file. You are unlikely to initially use most of the options in this menu; however, a few of them may be useful. For example, you can identify potential errors made when entering data by flagging possible duplicate entries using the **Identify Duplicate Cases** tool.
- **The 'Transform' menu (shortcut Alt + T)**: This menu allows you to manipulate your variables.
- **The 'Analyze' menu (shortcut Alt + A)**: This is the menu you will likely use the most and need first: Descriptive Statistics, Compare Means, General Linear Model, Correlation and Regression..., etc.
- **The 'Direct Marketing' menu (shortcut Alt + M)**: This is more for businesses wishing to conduct market research. You will not need to use this menu!
- **The 'Graphs' menu (shortcut Alt + G)**: This menu allows you to present data in graphical form, which will help you better understand your data. There are several ways to create graphs in SPSS, but this is a good starting point.
- **The 'Utilities' menu (shortcut Alt + U)**: In practice, it is useful for creating customized and automated analyses..., but feel free to ignore it for now!

- **The 'Window' menu (shortcut Alt + W)**: This menu allows you to quickly access other windows that might be hidden.
- **The 'Help' menu (shortcut Alt + H)**: This menu can be very useful, as it provides help and information both about the program system itself and the statistical tests it offers.

3. The toolbar: provides shortcuts to commonly used menu commands.
4. The Data Editor window: The name at the top of each column is the variable name, i.e., the name you will use to refer to a variable, while each row represents an observation (a case).
5. The tabs: Data View and Variable View. Data View is where we inspect our actual data, and Variable View is where we see additional information about our data.
6. The status bar: at the bottom of each window, SPSS provides several pieces of information such as command status, filter status, etc.

3.3 Data Manipulation in SPSS

■ SPSS data have three main components: observations, variables, and metadata. When you receive data, you will rarely have a problem with the observations, occasionally a problem with the variables, but almost always a problem with the metadata.

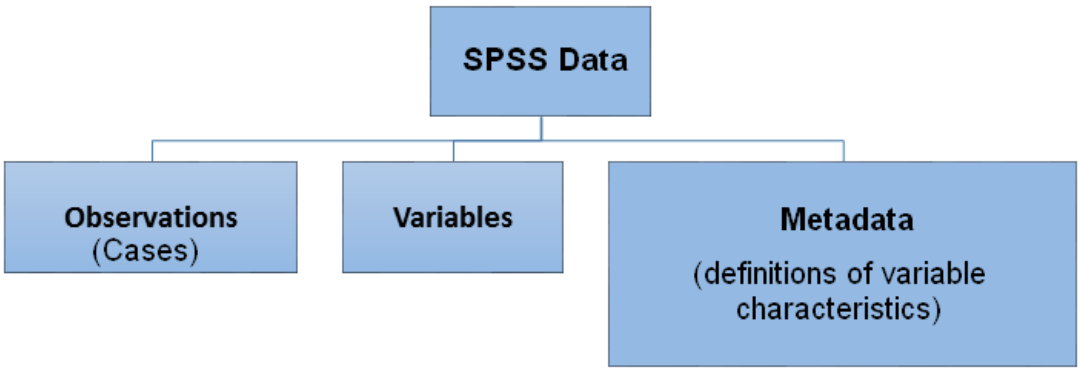


Figure 3.2: SPSS data.

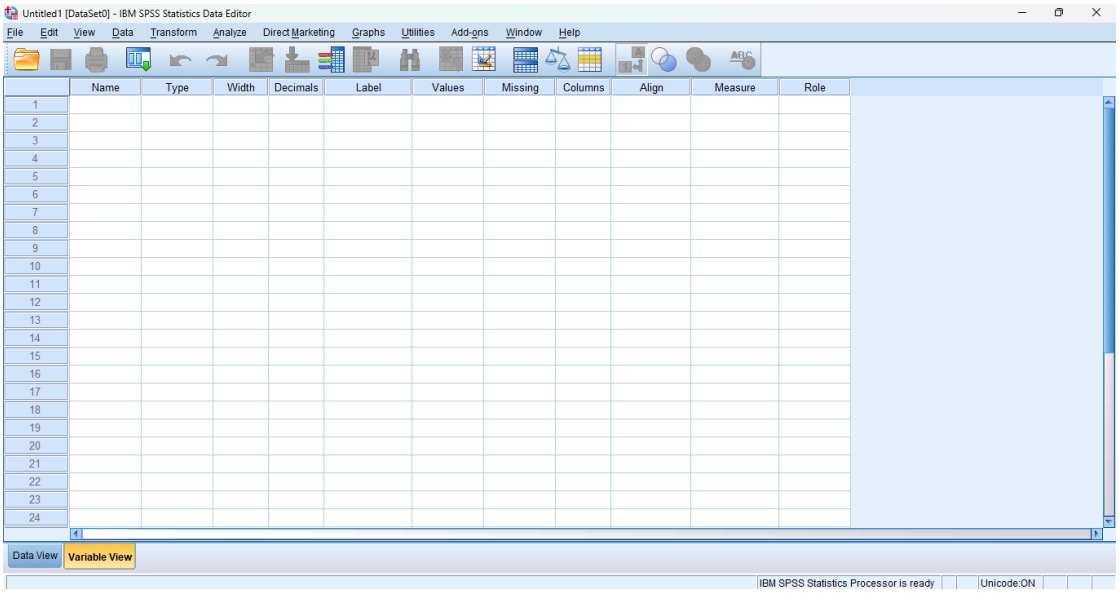


Figure 3.3: Variable View.

■ SPSS can read data from various formats, including databases, text files, Microsoft Excel, CSV... etc. You can also type directly into SPSS, and you can even paste copied data into SPSS.

3.3.1 Definition of Metadata

■ In SPSS, data are organized as observations (rows), and each observation consists of a set of variables (columns). First, you define the characteristics of the variables that make up an observation, then you enter the data into the variables that make up the contents of the observations.

■ To enter data in SPSS, use the "Variable View" tab. As you can see in the figure below (see Figure 3.3), the attributes of the variable (such as name, type, and width) are defined at the top of the window. All you have to do is enter something in each column for each variable.

■ The 11 characteristics are the only ones needed to fully specify all the attributes of a variable. When you add a new variable, you will notice that reasonable default values appear for most characteristics. The 11 characteristics of a variable are:

1. **Name**: Simply click on the cell and enter a short descriptive label, such as: Age, income, gender, patient.... Although you can enter longer names here, it is recommended to keep them short, as they will be used in named lists as well as for identification labels on data graphics and other formats where space may be limited.

    - Variable names must begin with a letter (A–Z or a–z).
    - They may contain letters, numbers (0–9), and underscore characters (_).

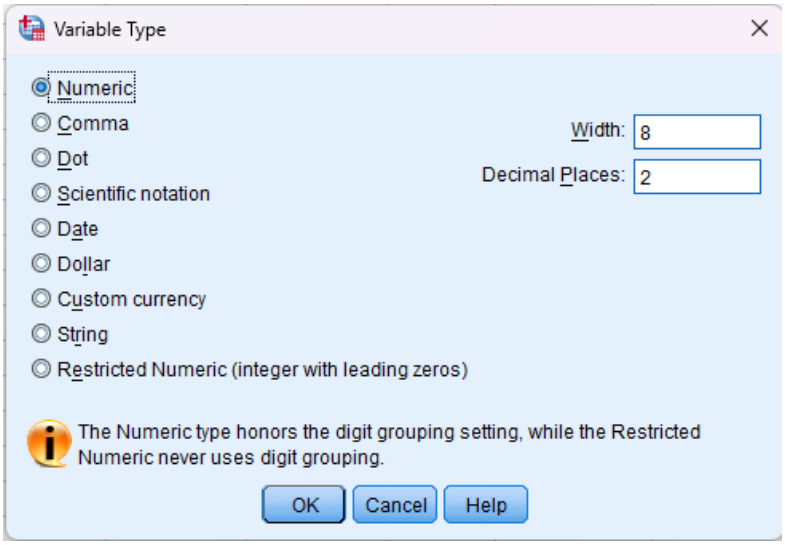


Figure 3.4: Variable type.

- No other special characters, spaces, or accented letters are allowed.
- The name must not be a reserved SPSS keyword, such as ALL, BY, AND, etc.
- Each variable name in the same data file must be unique.
- Variable names cannot begin with a number.

2. **Type**: Most of the data you enter will simply be regular numbers. However, data such as currency must be displayed in a special format, and data such as dates require special calculation procedures. For this type of data, you just need to specify the type you have, and SPSS takes care of the details for you. To display the dialog box shown in Figure 3.4, select a cell in the Type column, then click the button represented by three dots that appears.
3. **Width**: The Width column in the definition of a variable determines the number of characters used to display the value. If the value to display is not large enough to fill the space, the output will be padded with blanks. If it is larger than what you specified, it will be reformatted to fit or asterisks will be displayed.
4. **Decimals**: The Decimals column contains the number of digits that appear to the right of the decimal point when the value is displayed on the screen.
5. **Label**: The name and the label share the same fundamental purpose: they are descriptors that identify the variable. The difference is that the name is the short identifier, and the label is the long one. You may also choose to ignore defining the label. If you do not define a label for a variable, SPSS will use the variable name you defined for everything.

6. **Values**: The Values column is where you can assign labels to all possible values of a variable. To display the dialog box shown in Figure 3.5, select a cell in the Values column, then click the button represented by three dots that appears.

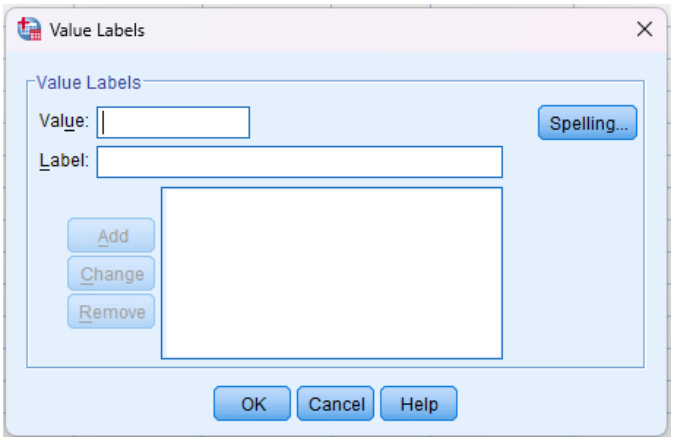


Figure 3.5: Value Labels.

Normally, you make a single entry for each possible value that a variable can take. But, for example, for a variable named Sex, you may assign the value 1 to the label Male and 2 to the label Female. If you define value labels, your output can display the labels instead of the numeric values.

To define a label for a value, follow these steps:

(a) In the Value box, enter the value.

(b) In the Label box, enter a label.

(c) Click the Add button. (The value and label will appear in the large text block.)

(d) To modify or delete a definition, simply select it in the text block, make your changes, then click the Change button.

(e) Repeat steps (a) to (d) as needed.

(f) To save the value labels and close the dialog box, click OK.

7. **Missing**: You can specify codes for missing data. To display the dialog box shown in Figure 3.6, select a cell in the Missing column. A button represented by three dots will appear—click it to open the dialog box.

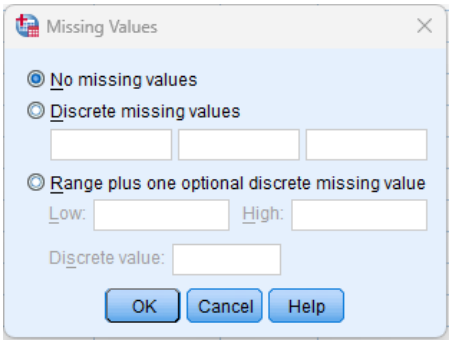


Figure 3.6: Missing values.

For example, suppose you are entering responses to questions, and one of the questions is: "How many children do you have?" The normal answer to this question is a number, so you define the variable type as numeric. You may define –99 as the value entered when the response is "I don't remember," and –98 when the response is "I cannot say."

8. **Columns**: In the Columns attribute, you can specify the width of the column used to enter the data.

9. **Align**: The Align column determines the position of the data within its allocated space.

10. **Measure**: Your value for the Measure attribute specifies the level of measurement of your variable. Here are the measurement level options in SPSS:

    - **Nominal**: A value that specifies a category or type of thing. For example, you may use 0 for Disapprove and 1 for Approve, or 1 to indicate Female and 2 to indicate Male.

    - **Ordinal**: A value that specifies the position (order) of something in a list. For example, first, second, and third are ordinal numbers.

    - **Scale**: A number that specifies a magnitude. The scale may represent distance, weight, age, or a count of something.

11. **Role**: You do not need to worry about the Role column for now.

3.4 Entering and Displaying Data Items in the "Data View" Tab

After defining the variables, you can begin entering the data. Click on the **Data View** tab in the Data Editor window. At the top of the columns, you will see the variable names. Entering data into any of these cells is simple: just click the cell and start typing.

3.5 Saving SPSS Data

All you need to do is choose **File ➜ Save or Save As (Ctrl + S)**, select your file type, and then enter a file name. The SPSS Statistics file format is "**.sav**" (see Figure 3.7).

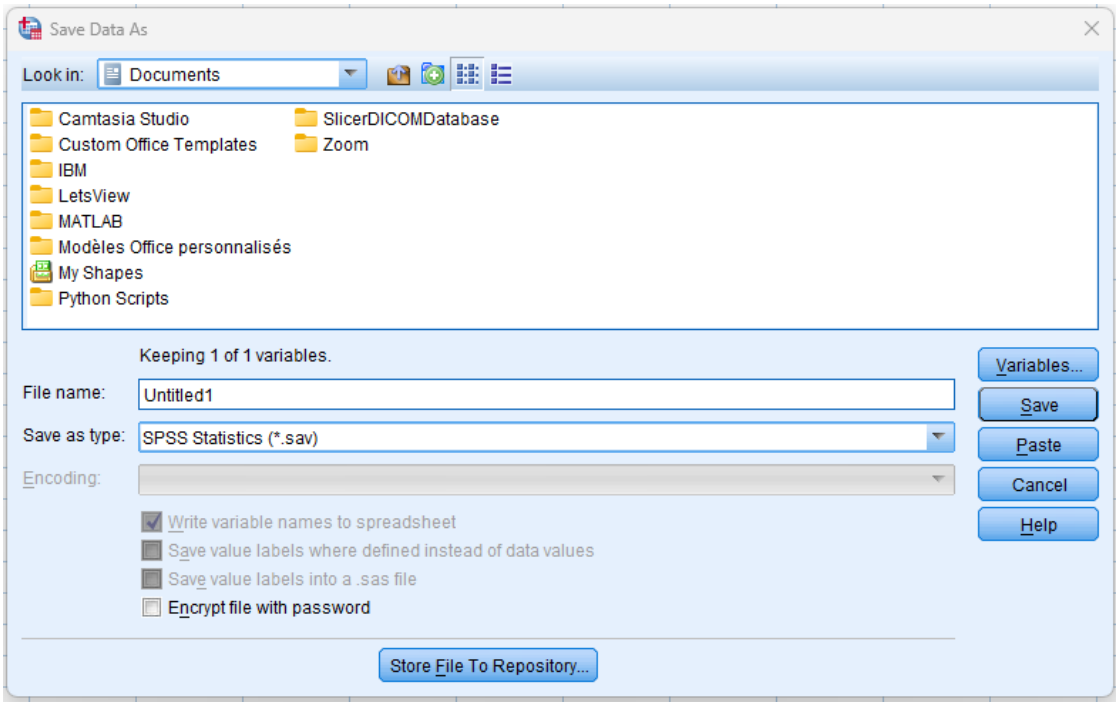


Figure 3.7: Saving Data in SPSS.

You have many file types to choose from: plain text formats, Excel spreadsheet formats, Lotus formats, dBase formats, SAS formats, SYLK format, Portable format, and 18 Stata formats.

3.6 Opening SPSS Data Files

To open a data file, choose **File ➜ Open ➜ Data (Ctrl + O)** and select the file to load. When you do this, the variable names and the data are loaded into SPSS.

3.7 Transferring Data from an Excel File to SPSS

To open your Excel file in SPSS:

1. **File ➜ Open ➜ Data (Ctrl + O)** from the SPSS menu.
2. Select the file type you want to open: Excel .xls, .xlsx, .xlsm.

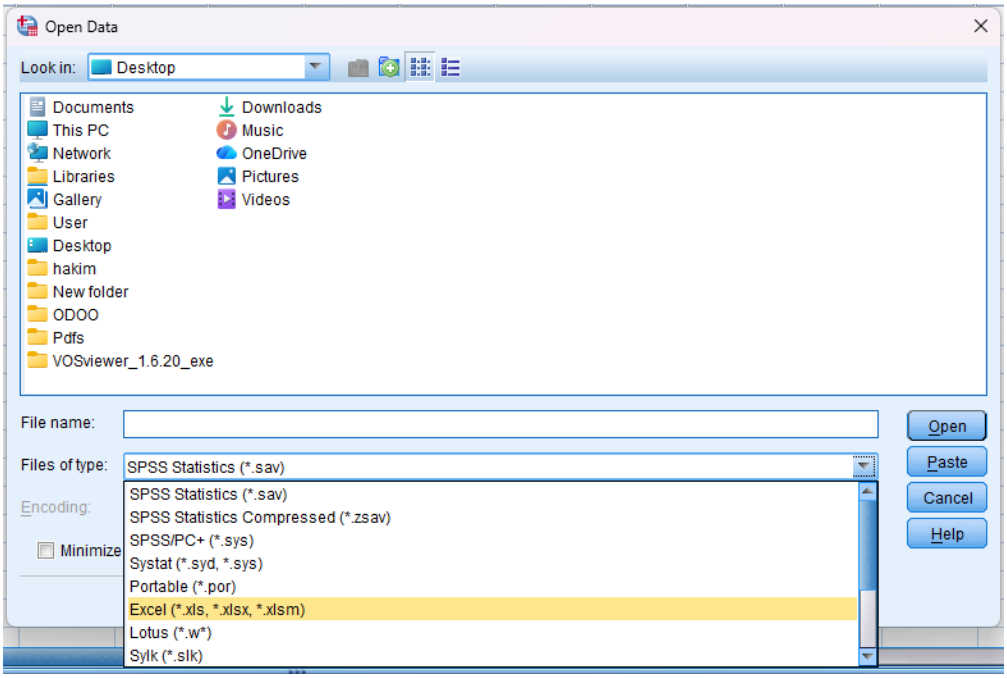


Figure 3.8: Opening Data in SPSS.

3. In the "Open Data" dialog box, select the file you want to open (see Figure 3.8).

4. Click **Open**.

5. The following dialog box appears (Figure 3.9).

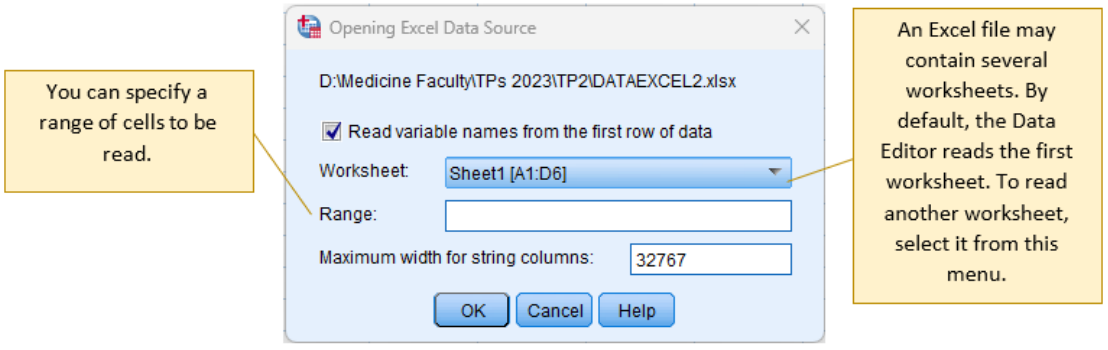


Figure 3.9: Open the Excel Data Source.

6. Click **OK**.

Note:
■ You can find the SPSS data file for this chapter at the following link: `https://aboulesnane.net/wp-content/datafiles/DataSPSS3.sav`

3.8 Conclusion

Organizing collected data in SPSS is important for accurate and efficient data management, research reproducibility, data security and confidentiality, and collaboration among researchers. SPSS offers a user-friendly interface and customizable options for organizing data, such as recoding, data transformations, and data cleaning. SPSS enables fast and efficient data management, including the ability to import and export data from various sources, merge datasets, and clean data. Overall, organizing collected data in SPSS is essential for conducting reliable and valid research in the social and biomedical sciences.

Chapter 4

Data Preprocessing

4.1 Introduction

Data preprocessing is a crucial step in data analysis that involves cleaning, transforming, and preparing raw data for analysis. It helps improve the quality, accuracy, and compatibility of data with analysis techniques and software tools. Data preprocessing can also reduce the time and effort required for data analysis by streamlining the data cleaning and transformation process, and it can help create data visualizations and summaries that facilitate the understanding and interpretation of the data.

4.2 Navigation in the SPSS Viewer

When you run an analysis, produce a chart, or even save a file, **the SPSS Statistics Viewer window** automatically appears to display what you have created (Figure 4.1).

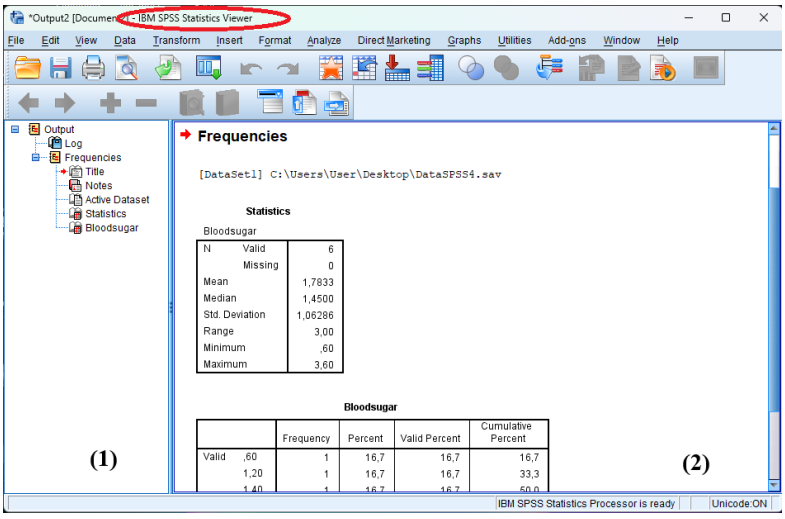


Figure 4.1: IBM SPSS statistics viewer.

■ The outline pane (1), on the left, contains an overview of all the information stored in the Viewer.
■ The content pane (2), on the right, contains statistical tables, charts, and text.
■ To hide an object in the content pane, you must:

✓ Select the object (table, chart, etc.) from the outline pane or the content pane.

✓ Click on the book icon in the toolbar  .

✓ To display this object again, click on the icon 

Hiding elements without deleting them allows the user to focus more easily on the results of interest while retaining all the results.

4.3 Working with Data in SPSS

	Patient	Weight	Bloodsugar	Sexe	Bloodgroup
1	P1	53,00	1,20	1	1
2	P2	-99,00	1,50	1	3
3	P3	88,00	1,40	2	2
4	P4	45,00	,60	1	4
5	P5	90,00	2,40	2	2
6	P6	175,00	3,60	2	1

Figure 4.2: SPSS data file.

1. Download the SPSS data file called "DataSPSS4.sav" from: `https://aboulesnane.net/wp-content/datafiles/DataSPSS4.sav`
2. The data contain five variables named: Patient, Weight, Bloodsugar, Sexe, and Bloodgroup (see Figure 4.2).
   (a) The variable "**Patient**" is a string-type variable.
   (b) The variable "**Weight**" is a continuous quantitative numeric variable. (For the Weight variable, missing values are represented by the number -99 (see Figure 3.6)).
   (c) The variable "**Bloodsugar**" is a continuous quantitative numeric variable.
   (d) The possible values for the qualitative variable "**Sexe**" are: 1=Homme and 2=Femme.
   (e) The possible values for the qualitative variable "**Bloodgroup**" are: 1=AB, 2=A, 3=B, and 4=O.

4.4 Replacing Missing Values

In order to replace missing values with acceptable values, we will use measures of central tendency such as the mean, the median, the mode, etc. For example, in the **Weight** variable of our data, we have a missing value represented by -99. In order not to disturb the distribution of our data, we can replace this value with the mean value of the series. To do this:

1. Choose **Transform ➜ Replace Missing Values**: the dialog box **Replace Missing Values**... appears.

2. Move the variable **Weight** to the "**New Variables**" box.

3. In the **Method** menu (see Figure 4.3), you can select the best method used to replace missing values (Linear Interpolation, Measures of Central Tendency, etc.). In our case, we will use **Series Mean**.

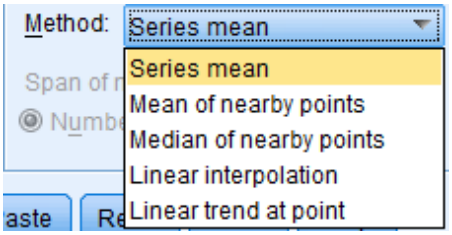


Figure 4.3: Method Menu.

4. Click on **OK** (see Figure 4.4).

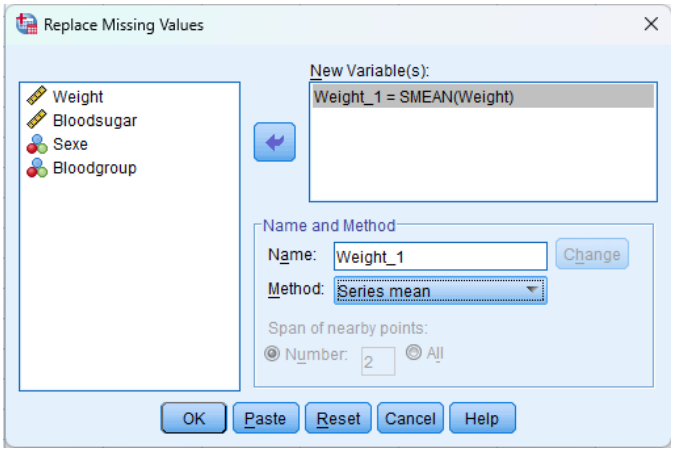


Figure 4.4: Replacing Missing Values.

As a result, a new column appears with the name "**Weight_1**", where the value -99 is replaced by the mean value 90.2 (see Figure 4.5).

	Patient	Weight	Bloodsugar	Sexe	Bloodgroup	Weight_1
1	P1	53,0	1,2	1	1	53,0
2	P2	-99,0	1,5	1	3	90,2
3	P3	88,0	1,4	2	2	88,0
4	P4	45,0	,6	1	4	45,0
5	P5	90,0	2,4	2	2	90,0
6	P6	175,0	3,6	2	1	175,0

Figure 4.5: Result after replacement.

4.5 Sorting Observations

1. Choose **Data ➜ Sort Cases**: the dialog box **Sort Cases** appears (Figure 4.6).

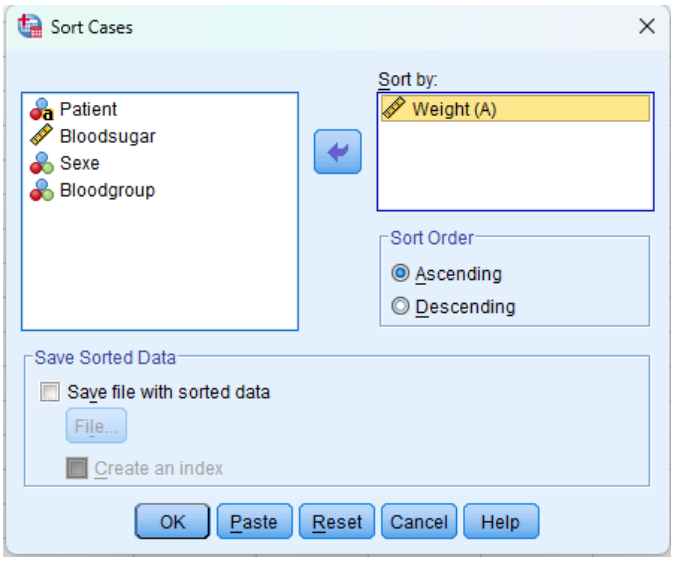


Figure 4.6: Sort Cases.

2. Select the variable **Weight**.
3. Click on **OK** to sort the data.

■ The order in which the data are displayed never affects the statistical analysis. You sort the data only to better view the information in the Data Editor.
■ It is possible to sort the data either in ascending order or in descending order. Note that by default, the data are sorted in **ascending** order.
■ Sorting can be based on one or more criteria.

Note:
■ To undo sorting:
✓ We cannot undo the sorting of the data.
✓ One tip is to sort the data again according to the "**Patient**" variable (repeat the same procedure described above, selecting the **Patient** variable instead of **Weight**).

4.6 Recoding Variables

By recoding variables, we can group a set of values into certain predefined categories depending on the type of data. This process helps save effort and time by restructuring the collected data in a better way. For example, suppose we want to recode the variable **Weight**, where we will group each set of values as follows:

1. Featherweight: [0-50]
2. Middleweight:]50-90]
3. Heavyweight:]90-120]
4. Super Heavyweight: >120

Here is how to do it in SPSS:

1. Choose **Transform ➜ Recode into Different Variables**: the dialog box **Recode into Different Variables** appears.
2. Click on the arrow button to move the variable **Weight** to the working area on the right.
3. Name the new output variable in the box on the right as **WeightCat ➜** Click on the **Change** button to save the new variable name, as shown in Figure 4.7.

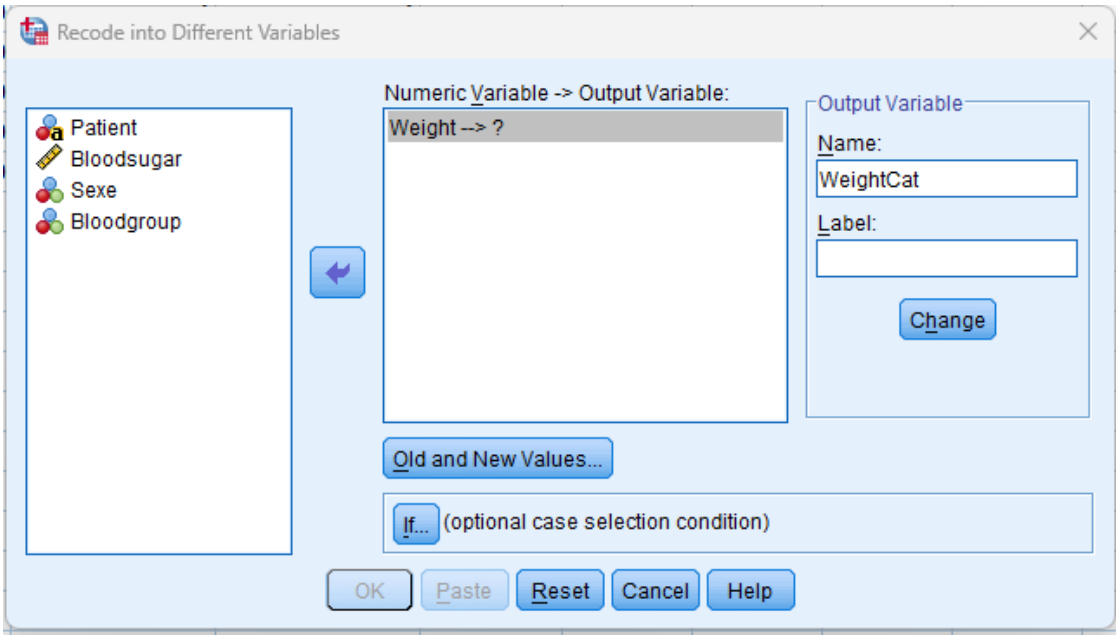


Figure 4.7: Dialog box: Recode into Different Variables.

4. Click on the **Old and New Values** button.
5. For categories 1, 2, and 3: select the **Range** radio button ➜ Enter the Min and Max values ➜ Next to the **Value** radio button: enter the category number ➜ Click on the **Add** button (see Figure 4.8).

6. For category 4: select the **Range, value through HIGHEST** radio button ➜ Enter the Min value (i.e., 120) ➜ Next to the **Value** radio button: enter the category number (i.e., 4) ➜ Click on the **Add** button (see Figure 4.8).

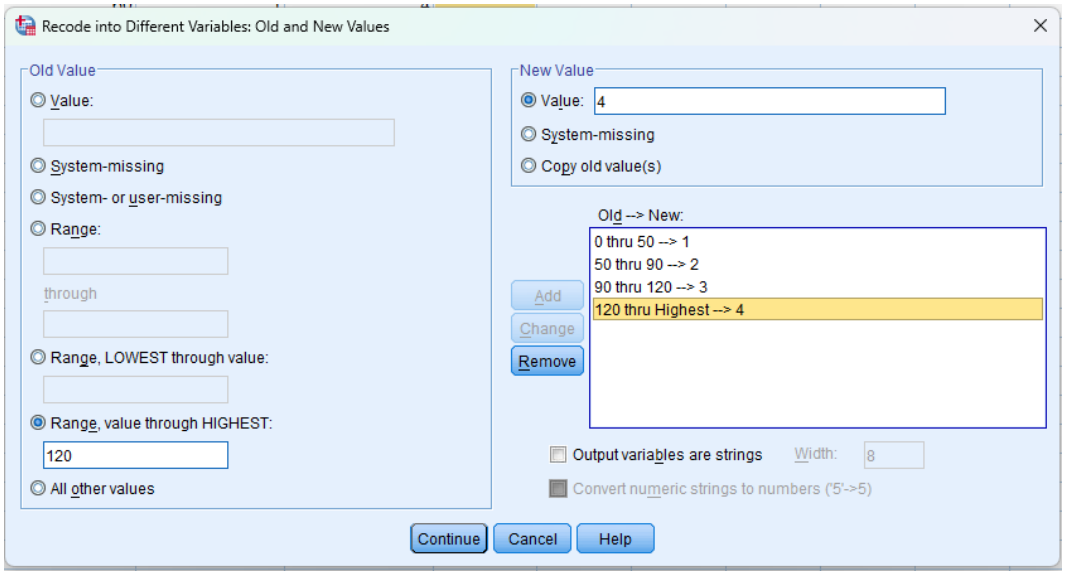


Figure 4.8: Dialog box: Old and New Values.

7. Click on **Continue**, then on **OK**. Figure 4.9 below shows the final result.

	Patient	Weight	Bloodsugar	Sexe	Bloodgroup	WeightCat
1	P1	53,00	1,20	1	1	2,00
2	P2	90,20	1,50	1	3	3,00
3	P3	88,00	1,40	2	2	2,00
4	P4	45,00	,60	1	4	1,00
5	P5	90,00	2,40	2	2	2,00
6	P6	175,00	3,60	2	1	4,00

Figure 4.9: Result after Recoding.

It is useful to tell anyone viewing your output what these recoded values represent. To do this, click on the **Variable View** tab at the bottom of the spreadsheet, then click in the Values field (for the new variable **WeightCat**) and add value labels as described previously (i.e., 1=Featherweight, 2=Middleweight, 3=Heavyweight, 4=Super Heavyweight).

4.7 Deleting a Variable or an Observation

➤ Suppose we want to delete the new variable **WeightCat**. To delete a variable in Data View, click on the variable name and press the **Delete** key on the keyboard, or right-click on the variable name and click **Clear**.

➤ To delete an observation (an entire row of data), follow these steps: Click on

the observation number on the left (the entire row will be highlighted), press **Delete** on the keyboard, or right-click on the observation number and click **Clear**.

4.8 Splitting Data

Sometimes, in comparative studies, we divide the data into several groups according to certain criteria and then perform our analyses on each subgroup separately. Data splitting allows this process to be carried out.

1. Choose **Data ➜ Split File**: the dialog box **Split File** appears.
2. Select the **Compare groups** radio button.
3. Choose **Sexe** as the comparison variable and click on **OK** (see Figure 4.10).

   Note: After this process, we notice that there is no significant change, except that the data have been reorganized according to the patients' sex.

4. Choose **Analyze ➜ Descriptive Statistics ➜ Frequencies**.
5. Choose **Bloodgroup** and place it in the Variable(s) box.
6. Click on **OK**: the resulting output, illustrated in Figure 4.11, is grouped by **Sexe**.

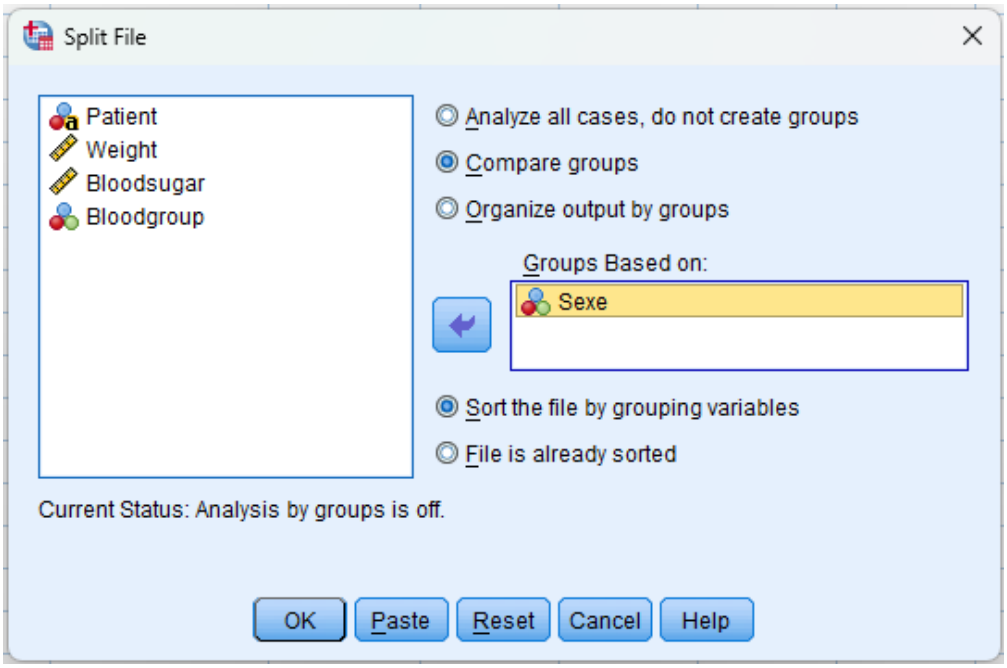


Figure 4.10: Split file.

Frequencies

Statistics

Bloodgroup

Homme	N	Valid	3
		Missing	0
Femme	N	Valid	3
		Missing	0

Bloodgroup

Sexe			Frequency	Percent	Valid Percent	Cumulative Percent
Homme	Valid	AB	1	33,3	33,3	33,3
		B	1	33,3	33,3	66,7
		O	1	33,3	33,3	100,0
		Total	3	100,0	100,0	
Femme	Valid	AB	1	33,3	33,3	33,3
		A	2	66,7	66,7	100,0
		Total	3	100,0	100,0	

Figure 4.11: Frequencies result.

Notes:
■ The split can be based on one or more criteria.
■ To cancel splitting, you must:

1. Choose **Data ➜ Split File**.
2. Select the **Analyze all cases, do not create groups** radio button, then click the **OK** button (or press the **Reset ➜ OK** button).

4.9 Data Selection

4.9.1 Simple Logical Condition

1. Choose **Data ➜ Select Cases**: the **Select Cases** dialog box appears, as illustrated in Figure 4.12.
2. Select the **If condition is satisfied** radio button, then click the **If**... button: You can now specify the selection criteria (see Figure 4.13).
3. Move the variable **Sexe** from the list on the left to the expression box (upper left): you can move the variable either by dragging it or by selecting it and then clicking the arrow button.
4. Use your keyboard or the on-screen numeric keypad to enter **=2** in the expression box.
5. Click on **Continue**, then on **OK**. The figure below shows the final result.

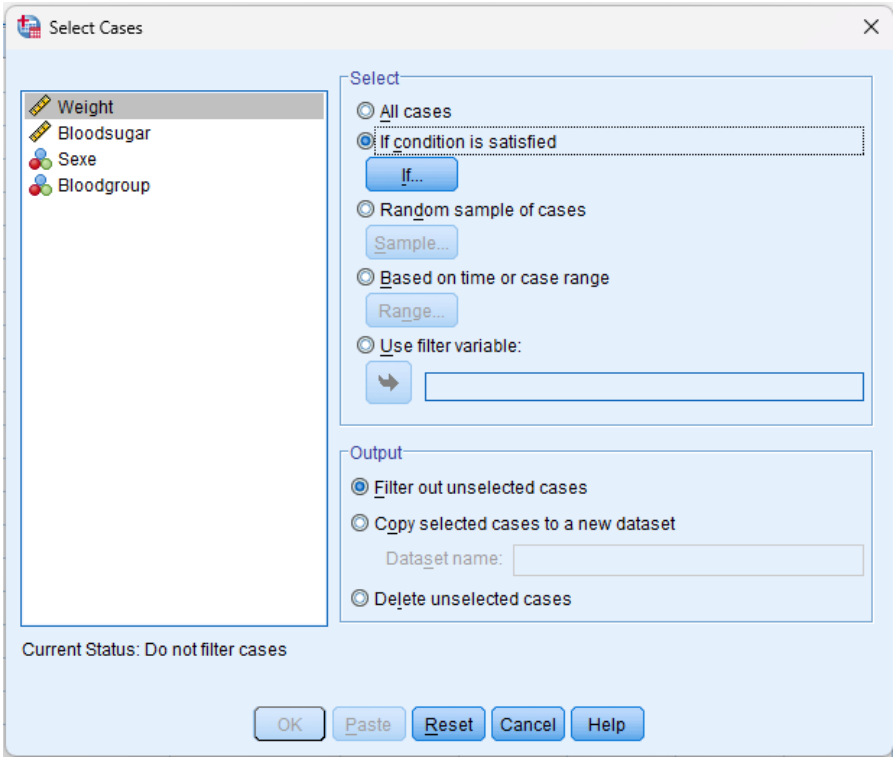


Figure 4.12: Select Cases.

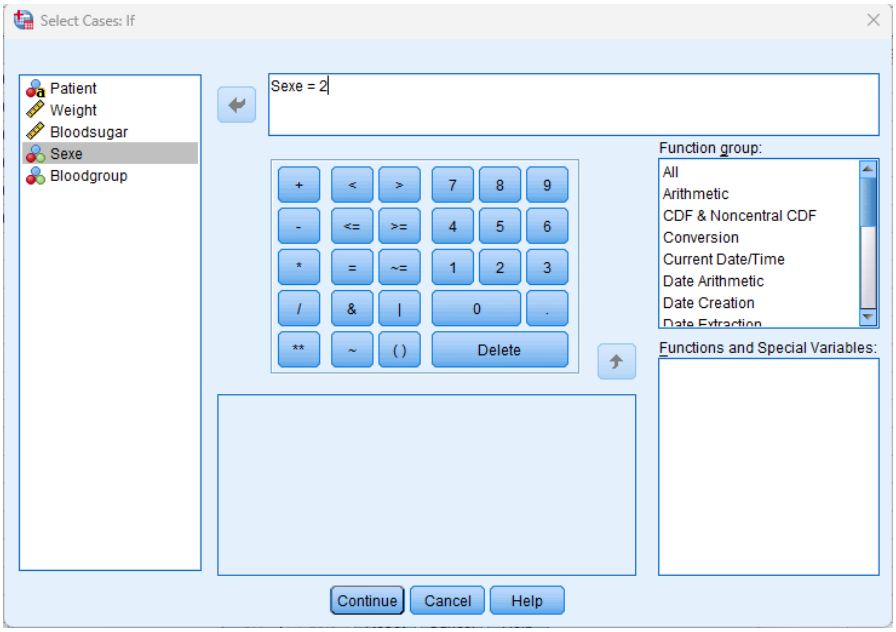


Figure 4.13: Conditional expression.

The diagonal slashes on some row IDs (in the first column) indicate that the patients (**Homme**) are ignored (for the moment) and that only female patients (**Femme**) are analyzed. The variable $filter\_\$$ is also created and contains 0 and 1 for non-selected cases and selected cases, respectively (see Figure 4.14).

	Patient	Weight	Bloodsugar	Sexe	Bloodgroup	filter_$
1	P1	53,00	1,20	1	1	0
2	P2	90,20	1,50	1	3	0
3	P3	88,00	1,40	2	2	1
4	P4	45,00	,60	1	4	0
5	P5	90,00	2,40	2	2	1
6	P6	175,00	3,60	2	1	1

Figure 4.14: Result after selection.

To see the effects of the selection:

(a) Choose **Analyze ➜ Descriptive Statistics ➜ Frequencies**.

(b) Choose **Bloodgroup** and place it in the Variable(s) box.

(c) Click on **OK**: the resulting output, illustrated in the figure below (Figure 4.15), shows that SPSS displays only the results of blood groups for female patients.

Frequencies

Statistics

Bloodgroup

N	Valid	3
	Missing	0

Bloodgroup

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	AB	1	33,3	33,3	33,3
	A	2	66,7	66,7	100,0
	Total	3	100,0	100,0	

Figure 4.15: Frequencies result.

4.9.2 Complex Logical Condition

Logical conditions are fundamental concepts in logic and programming, allowing decisions to be made based on the evaluation of certain expressions that result in true (True: T) or false (False: F).

Conjunction

Truth table of conjunction (**& : AND**) (T: True, F: False)

P	Q	P ∧ Q
T	T	T
T	F	F
F	T	F
F	F	F

Disjunction

Truth table of disjunction (**| : OR**) (T: True, F: False)

P	Q	P ∨ Q
T	T	T
T	F	T
F	T	T
F	F	F

Table 4.1: Logical conjunction and disjunction.

1. Logical Operators

The following logical operators allow combining or modifying statements:

- **AND** ($\wedge$, &) :
  - An expression is true if **all** conditions are true.
  - Example: $A \wedge B$ is true if A **and** B are true.
- **OR** ($\vee$, |) :
  - An expression is true if **at least one** condition is true.
  - Example: $A \vee B$ is true if A or B (or both) are true.
- **NOT** ($\neg$, $\sim$) :
  - Reverses a condition. True becomes false, and false becomes true.
  - Example: $\neg A$ is true if A is false.

2. Properties of Logical Operators

- **Commutative property** :

$$A \wedge B = B \wedge A$$
$$A \vee B = B \vee A$$

- **Associative property** :

$$(A \wedge B) \wedge C = A \wedge (B \wedge C)$$
$$(A \vee B) \vee C = A \vee (B \vee C)$$

- **Distributive property** :

$$A \wedge (B \vee C) = (A \wedge B) \vee (A \wedge C)$$

$$A \vee (B \wedge C) = (A \vee B) \wedge (A \vee C)$$

➤ Let us now try to select men who have a weight greater than or equal to 50 or patients who have blood group A.
➤ Enter the following condition: **(Sexe = 1 & Weight >= 50) | Bloodgroup = 2**, (see Figure 4.16) (Parentheses between conditions are very important).

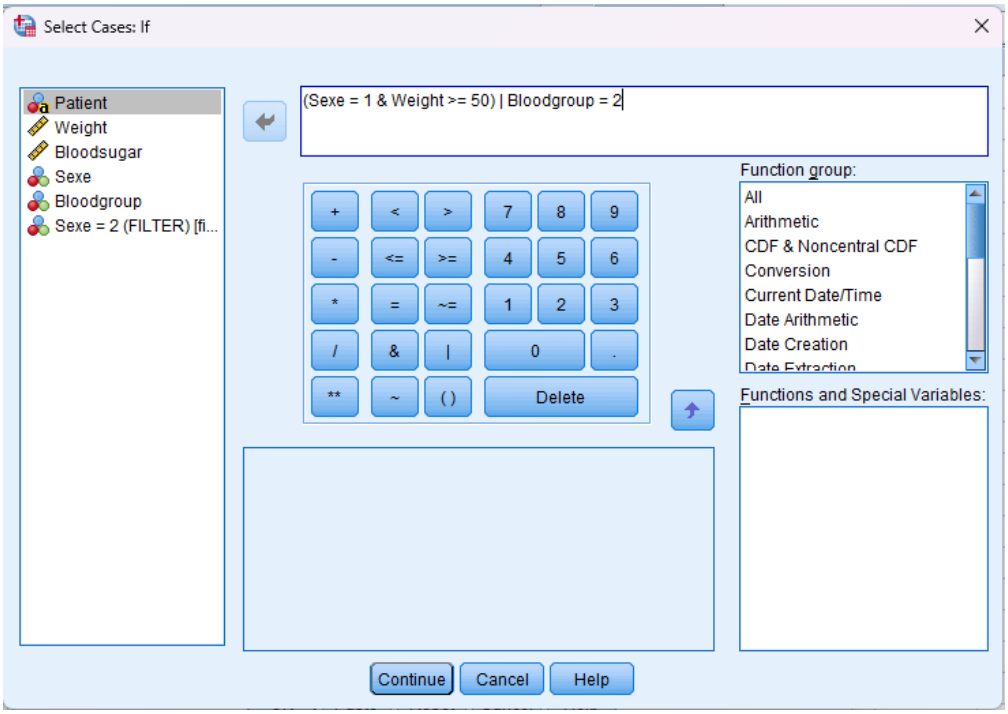


Figure 4.16: Conditional expression.

The result of the selection will be:

	Patient	Weight	Bloodsugar	Sexe	Bloodgroup	filter_$
1	P1	53,00	1,20	Homme	AB	Selected
2	P2	90,20	1,50	Homme	B	Selected
3	P3	88,00	1,40	Femme	A	Selected
4	P4	45,00	,60	Homme	O	Not Selected
5	P5	90,00	2,40	Femme	A	Selected
6	P6	175,00	3,60	Femme	AB	Not Selected

Figure 4.17: Result after selection.

Notes:

■ To cancel the selection, you must:

1. Choose **Data ➜ Select Cases**.
2. Select the **All cases** radio button, then click the **OK** button (or press the **Reset ➜ OK** button).

■ The selection can be based on one or more criteria.

4.10 Conclusion

Data preprocessing is a fundamental step in data analysis, as it directly affects the reliability and validity of the results obtained. Throughout this chapter, we have explored several essential preprocessing operations in SPSS, including handling missing values, sorting and recoding variables, selecting and splitting data, and applying logical conditions. These techniques allow analysts to clean, structure, and organize data efficiently before performing statistical analyses. By using SPSS preprocessing tools appropriately, users can improve data quality, ensure better compatibility with analytical methods, and extract more accurate and meaningful insights from their data.

Chapter 5

Data Analysis

5.1 Introduction

In biomedical data analysis, qualitative and continuous variables are often analyzed together to better understand complex phenomena, such as the risk factors of a particular disease or the effectiveness of a medical treatment. Therefore, it is important to have a thorough understanding of statistical methods for both types of variables in order to accurately interpret biomedical data and make informed decisions.

5.2 Data Collection in SPSS

	Patient	Weight	Bloodsugar	Sexe	Bloodgroup
1	P1	63,00	1,20	1	1
2	P2	90,20	1,70	1	3
3	P3	88,00	1,40	2	2
4	P4	45,00	,60	1	4
5	P5	90,00	2,40	2	2
6	P6	175,00	3,60	2	1
7	P7	63,00	2,10	1	2

Figure 5.1: SPSS Data File.

1. Download the SPSS data file called "**DataSPSS5.sav**" from: `https://aboulesnane.net/wp-content/datafiles/DataSPSS5.sav`
2. The data contain five variables named: Patient, Weight, Bloodsugar, Sexe, and Bloodgroup (see Figure 5.1).
   (a) The variable "**Patient**" is a **String** type variable.
   (b) The variable "**Weight**" is a continuous quantitative variable of Numeric type.

(c) The variable "**Bloodsugar**" is a continuous quantitative variable of Numeric type.

(d) The possible values for the qualitative variable "**Sexe**" are: 1=Homme and 2=Femme.

(e) The possible values for the qualitative variable "**Bloodgroup**" are: 1=AB, 2=A, 3=B, and 4=O.

5.3 Data Preprocessing in SPSS

■ The patient data were sorted according to their blood sugar (**Bloodsugar**) and weight (**Weight**) values. (see section 4.5).

The result of the sorting is:

	Patient	Weight	Bloodsugar	Sexe	Bloodgroup
1	P4	45,00	,60	1	4
2	P7	60,00	1,20	1	2
3	P1	63,00	1,20	1	1
4	P3	88,00	1,40	2	2
5	P2	90,20	1,70	1	3
6	P5	90,00	2,40	2	2
7	P6	175,00	3,60	2	1

Figure 5.2: Result after sorting.

■ The patient data were split according to their blood group (**Bloodgroup**) values. (see section 4.8).

The result of the splitting is:

	Patient	Weight	Bloodsugar	Sexe	Bloodgroup
1	P1	63,00	1,20	1	1
2	P6	175,00	3,60	2	1
3	P7	60,00	1,20	1	2
4	P3	88,00	1,40	2	2
5	P5	90,00	2,40	2	2
6	P2	90,20	1,70	1	3
7	P4	45,00	,60	1	4

Figure 5.3: Result after splitting.

5.4 Use of Descriptive Statistics

5.4.1 Frequencies for categorical (qualitative) variables

■ The most common technique for describing categorical data—nominal and ordinal levels of measurement—is to request **a frequency table**, which provides a summary indicating the number and percentage of cases falling into each category of a variable. Users can also request additional summary statistics such as the mode or the median, among others.
■ Here is how to run the Frequencies procedure in order to create a frequency table that will allow you to obtain summary statistics for qualitative variables:

1. Choose **Analyze ➜ Descriptive Statistics ➜ Frequencies**: The Frequencies dialog box appears. In this example, and based on the **previous data split**, you want to study the distribution of the variable **Sexe** (Homme, Femme) for each value of **Bloodgroup** (AB, A, B, O).

2. Select the variable **Sexe**, and place it in the Variable(s) box, as illustrated in Figure 5.4.

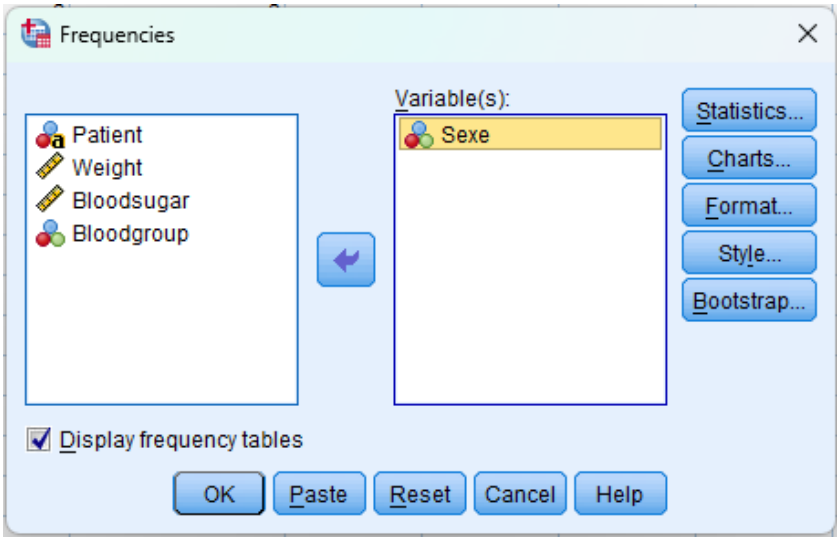


Figure 5.4: Frequencies dialog box.

3. Click the **Statistics** button: the **Frequencies: Statistics** dialog box is displayed (see Figure 5.5).

4. In the **Central Tendency** section, check the **Mode** box, as illustrated in the figure below. This dialog box provides many statistics, but it is essential that you request only those that correspond to the level of measurement of the variables that you placed in the Variable(s) box. For nominal variables, the most appropriate measure of central tendency is the **mode**, because these variables have neither order nor hierarchy. For ordinal variables, the **median** is preferred, since it takes into account the order of the categories, but the mode can also be used if one wishes to identify the most frequent category.

5. Click **Continue**.

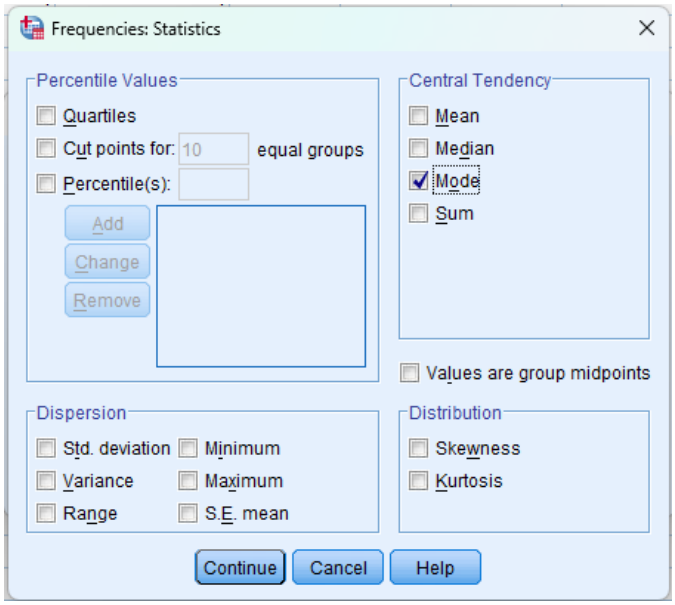


Figure 5.5: Frequencies: Statistics dialog box.

6. Click the **Charts** button: the **Frequencies: Charts** dialog box is displayed.
7. In the **Chart Type** section, select the **Bar charts** radio button; in the Chart Values section, select the **Percentages** radio button (see Figure 5.6).

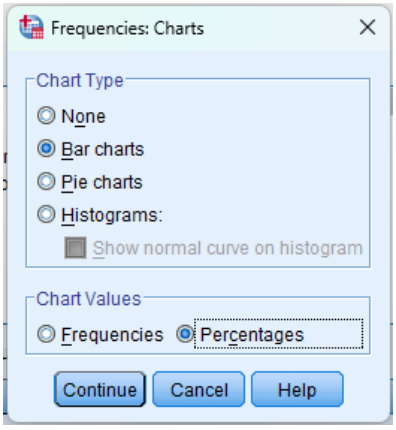


Figure 5.6: Frequencies: Charts dialog box.

8. Click **Continue**, then **OK**: SPSS runs the Frequencies procedure and calculates the summary statistics, the frequency table, and the bar chart that you requested.
9. The resulting output, illustrated in the figures below, is grouped by **Bloodgroup** (see Figure 5.7).

Frequencies

[Jeu_de_données1] C:\Users\User\Desktop\DataSPSS5.sav

Statistics

Sexe

AB	N	Valid	2
		Missing	0
	Mode		1[a]
A	N	Valid	3
		Missing	0
	Mode		2
B	N	Valid	1
		Missing	0
	Mode		1
O	N	Valid	1
		Missing	0
	Mode		1

a. Multiple modes exist. The smallest value is shown

(Effectif) Absolute frequencies

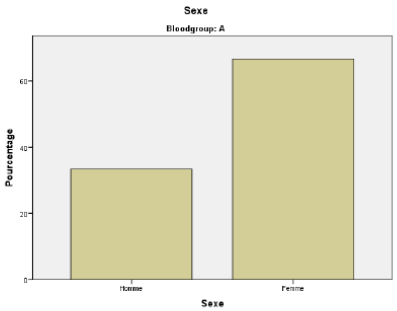
(Fréquences) Relative frequencies (%)

Cumulative frequencies (%)

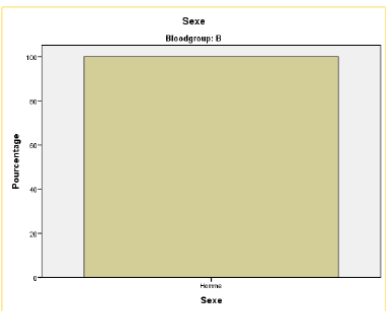
Sexe

Bloodgroup			Frequency	Percent	Valid Percent	Cumulative Percent
AB	Valid	Homme	1	50,0	50,0	50,0
		Femme	1	50,0	50,0	100,0
		Total	2	100,0	100,0	
A	Valid	Homme	1	33,3	33,3	33,3
		Femme	2	66,7	66,7	100,0
		Total	3	100,0	100,0	
B	Valid	Homme	1	100,0	100,0	100,0
O	Valid	Homme	1	100,0	100,0	100,0

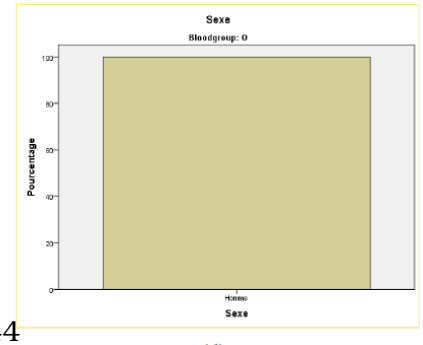
(a)



(b)



(c)



(d)

Figure 5.7: Analysis result.

Note:
In IBM SPSS Statistics Viewer, you can change the style of the plots by using the Chart Editor:

1. Double-click the chart in the Statistics Viewer to open the Chart Editor.
2. Once the **Chart Editor** is open, you can change the style of elements by double-clicking the element you want to modify (for example, a bar, an axis, or a legend). This will open a **Properties** window where you can adjust the style, color, size, and other settings.

5.4.2 Frequencies for continuous variables

■ As you have seen, **frequency tables** display counts and percentages, which is extremely useful when working with qualitative variables. However, for continuous variables, which can have many values, frequency tables become less useful.
■ To run frequencies for continuous variables, proceed as follows:

1. **Cancel the data split**.
2. Choose **Analyze ➜ Descriptive Statistics ➜ Frequencies**.
3. Select the variables **Weight** and **Bloodsugar**, and place them in the Variable(s) box.
4. Uncheck the **Display frequency tables** option, as illustrated in Figure 5.8 below.

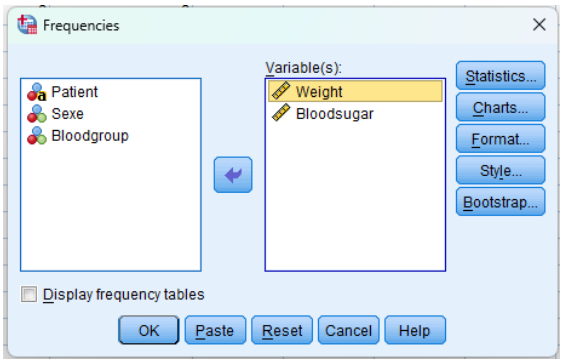


Figure 5.8: Frequencies dialog box.

5. Click the **Statistics** button.
6. In the **Central Tendency** section, check the boxes **Mean, Median**, and **Mode**. In the **Dispersion** section, select **Standard Deviation, Variance, Minimum, and Maximum**. You can select more metrics like : Quartiles, Sum... etc.
7. Click **Continue**.
8. Click the **Charts** button.
9. Select the **Histograms** radio button and check the **Display normal curve on histogram** option, as illustrated in Figure 5.9.

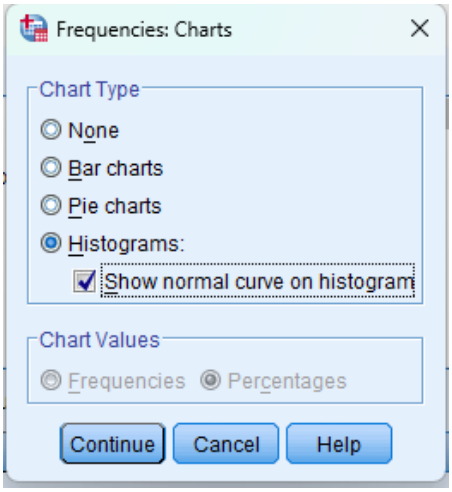


Figure 5.9: Frequencies: Charts dialog box.

Frequencies

Statistics

		Weight	Bloodsugar
N	Valid	7	7
	Missing	0	0
Mean		87,3143	1,7286
Median		88,0000	1,4000
Mode		45,00[a]	1,20
Std. Deviation		42,49029	,99115
Variance		1805,425	,982
Minimum		45,00	,60
Maximum		175,00	3,60
Sum		611,20	12,10
Percentiles	25	60,0000	1,2000
	50	88,0000	1,4000
	75	90,2000	2,4000

a. Multiple modes exist. The smallest value is shown

Histogramme

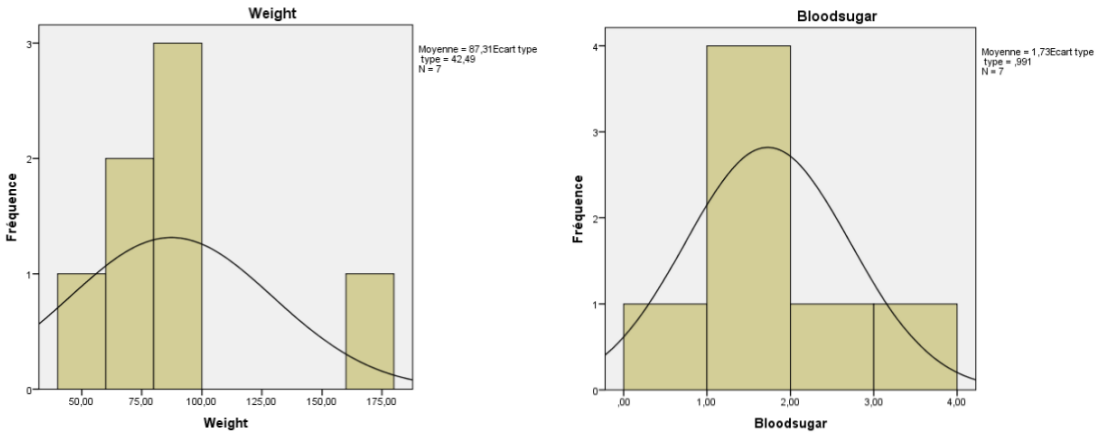


Figure 5.10: Analysis result.

10. Click **Continue**, then **OK**: SPSS runs the Frequencies procedure and calculates the summary statistics and the histogram that you requested.
11. The resulting output is illustrated in Figure 5.10.

5.4.3 Summarizing continuous variables with the Descriptives procedure

■ The Descriptives procedure provides a concise summary of various statistics and the number of cases with valid values for each variable included in the table.
■ To use the Descriptives procedure, proceed as follows:

1. Choose **Analyze ➜ Descriptive Statistics ➜ Descriptives**: the Descriptives dialog box appears.
2. Select the variables **Weight** and **Bloodsugar**, and place them in the Variable(s) box.
3. Click **OK**: SPSS runs the Descriptives procedure and calculates the summary statistics as shown below (Figure 5.11).

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Weight	7	45,00	175,00	87,3143	42,49029
Bloodsugar	7	,60	3,60	1,7286	,99115
Valid N (listwise)	7				

Figure 5.11: Descriptive statistics result.

5.5 Conclusion

In this chapter, we explored different methods for analyzing biomedical data using SPSS, with a focus on qualitative and continuous variables. These approaches made it possible to summarize the data in a systematic way, while adapting to the specific characteristics of the variables studied.
The use of descriptive statistics illustrates the power of SPSS to organize and analyze biomedical data efficiently. By highlighting the importance of a rigorous and appropriate methodology, this chapter provides a solid foundation for conducting exploratory analyses, which are essential for identifying trends, correlations, or significant differences.

Chapter 6

Analysis of Relationships Between Statistical Variables

6.1 Introduction

Biomedical data analysis can be used to study the relationships between variables in various contexts, such as investigating the risk factors of a disease or evaluating the effectiveness of a medical treatment. The analysis of relationships between variables involves examining the association or correlation between different variables, which may be categorical or continuous. Statistical methods such as cross-tabulation and regression analysis can be used to identify and quantify relationships between qualitative and continuous variables, respectively. In addition, correlation analysis can be used to determine the strength and direction of the relationship between two continuous variables. By understanding the relationships between variables, researchers can better comprehend the underlying factors that contribute to a particular outcome or phenomenon and make informed decisions based on their analysis.

6.2 Data Collection in SPSS

	Patient	Weight	Bloodsugar	LungCancer	Smoking
1	P1	63,00	1,20	1	2
2	P2	90,20	1,70	1	3
3	P3	88,00	1,40	0	2
4	P4	45,00	,60	1	3
5	P5	90,00	2,40	0	2
6	P6	175,00	3,60	0	1
7	P7	60,00	1,20	1	2
8	P8	120,00	1,92	0	1
9	P9	55,00	,70	0	1
10	P10	160,00	4,62	1	3

Figure 6.1: SPSS Data File.

1. Download the SPSS data file named "**DataSPSS6.sav**" from: `https://aboulesnane.net/wp-content/datafiles/DataSPSS6.sav`
2. The data contain five variables named: Patient, Weight, Bloodsugar, LungCancer, and Smoking (see Figure 6.1).
   (a) The variable "**Patient**" is a **String** variable.
   (b) The variable "**Weight**" is a continuous quantitative variable of Numeric type.
   (c) The variable "**Bloodsugar**" is a continuous quantitative variable of Numeric type.
   (d) The possible values for the qualitative variable "**LungCancer**" are: 0=Non et 1=Oui.
   (e) The possible values for the qualitative variable "**Smoking**" are: 1=Non , 2=Parfois, 3=Beaucoup.

6.3 Data Preprocessing in SPSS

■ The patient data were sorted according to their Weight values (**Weight**). (see Section 4.5).

The result of the sorting is:

	Patient	Weight	Bloodsugar	LungCancer	Smoking
1	P4	45,00	,60	1	3
2	P9	55,00	,70	0	1
3	P7	60,00	1,20	1	2
4	P1	63,00	1,20	1	2
5	P3	88,00	1,40	0	2
6	P5	90,00	2,40	0	2
7	P2	90,20	1,70	1	3
8	P8	120,00	1,92	0	1
9	P10	160,00	4,62	1	3
10	P6	175,00	3,60	0	1

Figure 6.2: Result after sorting.

6.4 Bivariate Statistical Distributions

6.4.1 Relationships between Categorical (Qualitative) Variables

One of the most common ways to analyze data is by using cross-tabulations. As mentioned, you use a cross-tabulation when you want to study the relation-

		Independent Variables	
	Variables	**Qualitative**	**Quantitative**
Dependent Variables	**Qualitative**	Cross-tabulations, Nonparametric Tests	Logistic Regression, Discriminant Analysis
	Quantitative	T-Test, ANOVA	Correlation, Linear Regression

Table 6.1: Analysis of Statistical Relationships.

ship between two or more categorical variables. For example, you may wish to examine the impact (relationship) of cigarette smoking on lung cancer.

Here is how to perform a cross-tabulation using our data (between the variables **LungCancer** and **Smoking**) :

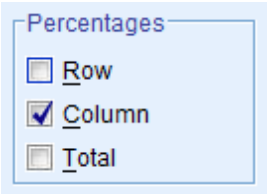
1. Choose **Analyze ➜ Descriptive Statistics ➜ Crosstabs** : The Crosstabs dialog box appears.
   Although you can place variables in either the Rows or Columns areas, **it is customary to place the independent variable in the column of the cross-tabulation table**. In bivariate analyses, the independent variable is the one that, in theory, influences the other variable, called the dependent variable.

   The independent variable in this study is smoking behavior, as it is assumed that smoking has a direct impact on the development of lung cancer:

2. Select the variable **LungCancer** and move it into the Row(s) box.

3. Select the variable **Smoking** and move it into the Column(s) box, as illustrated in the figure below (Figure 6.3).

4. Click on the **Cells** button: select row percentages, column percentages, or both.



5. Click Continue, then click **OK**.

6. The resulting output is illustrated in the figures below (Figure 6.4).

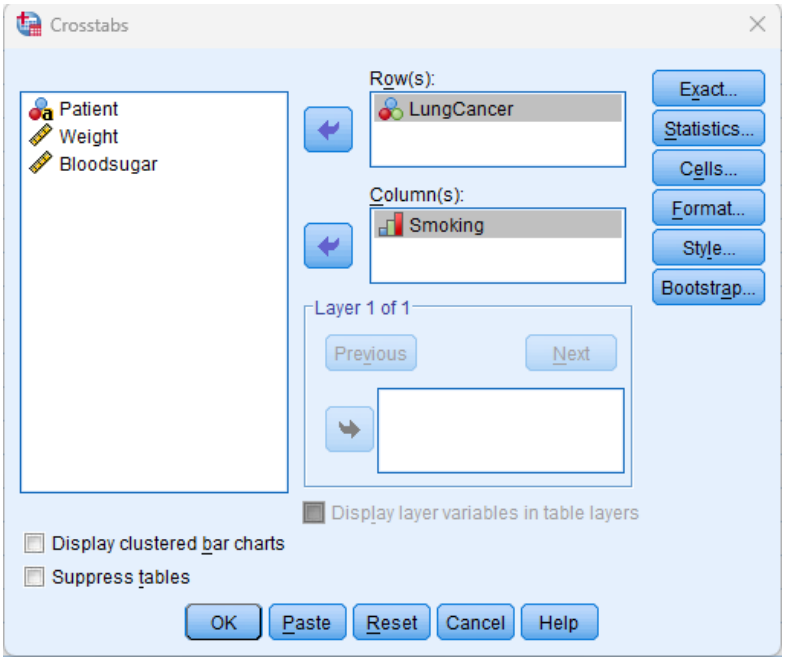


Figure 6.3: Crosstabs Dialog Box.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
LungCancer * Smoking	10	100,0%	0	0,0%	10	100,0%

LungCancer * Smoking Crosstabulation

			Smoking			Total
			Non	Parfois	Beaucoup	
LungCancer	Non	Count	3	2	0	5
		% within Smoking	100,0%	50,0%	0,0%	50,0%
	Oui	Count	0	2	3	5
		% within Smoking	0,0%	50,0%	100,0%	50,0%
Total		Count	3	4	3	10
		% within Smoking	100,0%	100,0%	100,0%	100,0%

Figure 6.4: Crosstabs Output.

Attention:
Correlation does not imply causation: means that just because two variables (like ice cream sales: High/Low, and shark attacks: Yes/No) change together, it does not mean one directly causes the other. They may be linked by a third, hidden factor (hot weather or season: Summer/Winter), coincidental timing, or random chance.

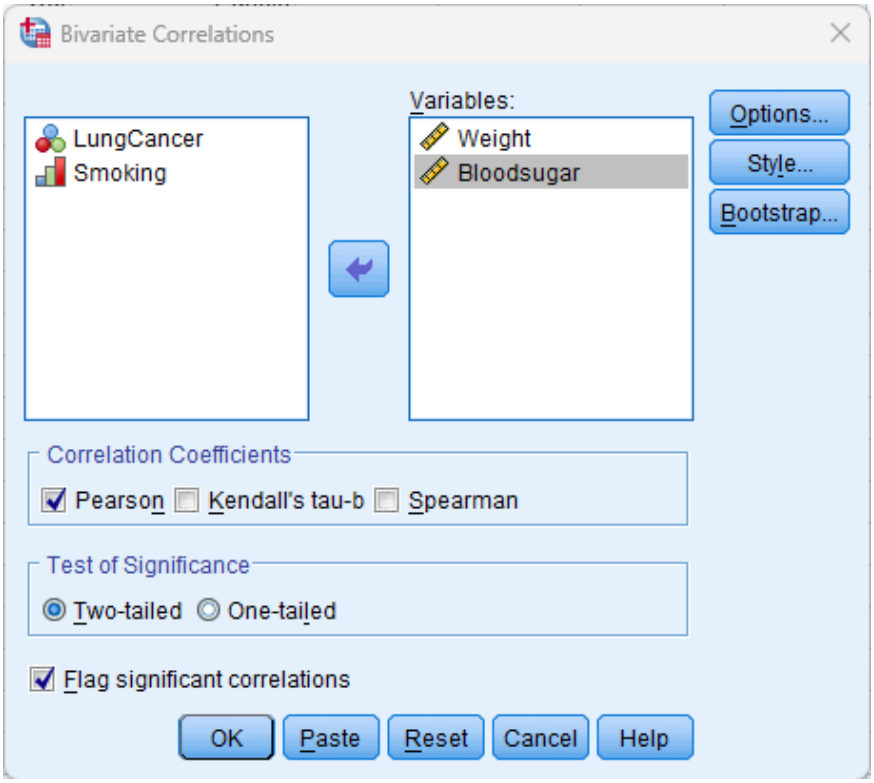


Figure 6.5: Bivariate Correlations.

6.4.2 Relationships between Quantitative Variables

The two statistical techniques most commonly used to analyze relationships between continuous variables are Pearson correlation and linear regression. Many people use the term correlation to refer to the idea of a relationship between variables in a model. In other words, variables are correlated with each other because changes in one variable affect the other.
While correlation simply attempts to determine whether two variables are related, linear regression goes a step further and attempts to predict the values of one variable based on another variable.

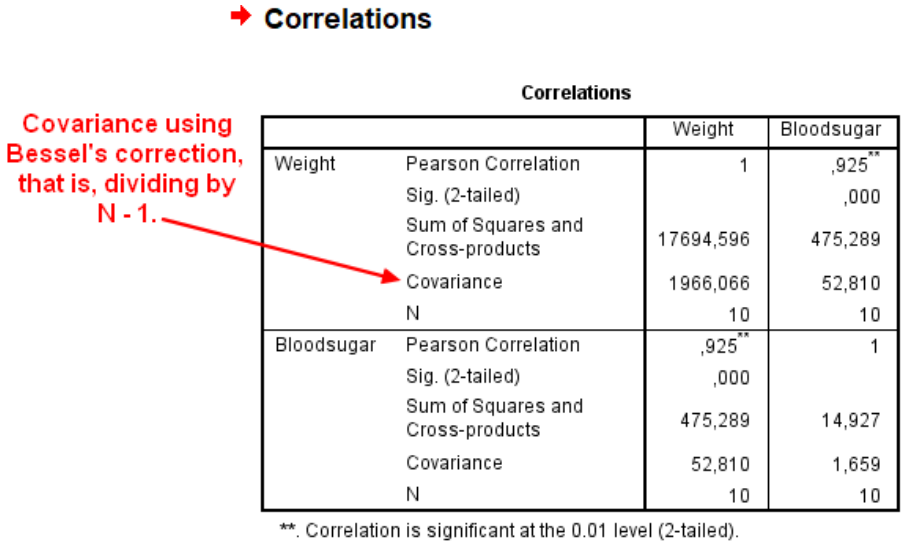
❖ Running the Bivariate Procedure

The Pearson correlation coefficient is a measure of the extent to which there is a linear (straight-line) relationship between two variables. It takes values between -1 and $+1$, so the larger the absolute value, the stronger the correlation. For example, a correlation of $+1$ indicates that the data fall on a perfectly upward-sloping straight line (positive relationship), and a correlation of -1 represents data forming a perfectly downward-sloping straight line (negative relationship). A correlation of 0 indicates that no linear relationship exists.
To test a correlation, proceed as follows:

1. Choose **Analyze ➜ Correlate ➜ Bivariate**. The **Bivariate Correlations** dialog box appears.
   In this example, you will study whether blood sugar level is related to weight. Note that there is no designation of dependent and independent

variables. Correlations will be calculated for all pairs of variables.

2. Select the variables **Weight** and **Bloodsugar** and place them into the Variables box, as illustrated in Figure 6.5 above.
3. Click on the **Options** button and check the option "**Cross-product deviations and covariances**".
4. Click **Continue**, then click **OK**.
5. The resulting output, illustrated in Figure 6.6 below.



Correlations

		Weight	Bloodsugar
Weight	Pearson Correlation	1	,925**
	Sig. (2-tailed)		,000
	Sum of Squares and Cross-products	17694,596	475,289
	Covariance	1966,066	52,810
	N	10	10
Bloodsugar	Pearson Correlation	,925**	1
	Sig. (2-tailed)	,000	
	Sum of Squares and Cross-products	475,289	14,927
	Covariance	52,810	1,659
	N	10	10

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 6.6: Correlation Table

In this example, you have a very strong positive correlation (0.925) that is statistically significant between bloodsugar and weight.
In another way, from the same table, we can calculate the correlation coefficient from the covariance matrix, as follows:

$$r = \frac{XY covariance}{\sqrt{X variance}\sqrt{Y variance}}$$

$$r = \frac{\left(\frac{\sum(X-\bar{X})(Y-\bar{Y})}{N}\right)}{\sqrt{\frac{\sum(X-\bar{X})(X-\bar{X})}{N}} * \sqrt{\frac{\sum(Y-\bar{Y})(Y-\bar{Y})}{N}}}$$

$$r = \frac{\left(\frac{475{,}289}{10}\right)}{\sqrt{\frac{17694{,}596}{10}} * \sqrt{\frac{14{,}927}{10}}} = 0.925$$

❖ Running the Simple Linear Regression Procedure

Correlations allow you to determine whether two continuous variables are linearly related to each other. Regression analysis consists of predicting the future (the unknown) based on data collected in the past (the known). Regression allows you to further quantify relationships by developing an **equation** so that you can predict, for example, blood sugar level based on the patient's body weight.

Linear regression is a statistical technique used to predict a **continuous dependent** variable from one or more **continuous independent** variables.

Since we have a strong linear correlation between blood sugar and weight, we can perform a linear regression, as follows:

1. Select **Analyze ➜ Regression ➜ Linear**.
   The Linear Regression dialog box appears. In this example, you want to predict blood sugar level from weight. You can place the dependent variable, blood sugar (**Bloodsugar**), into the **Dependent** box; this is the variable for which you want to define a prediction equation. You can place the predictor variable **Weight** into the **Independent(s)** box; this is the variable you will use to predict the dependent variable.
2. Select the variable **Bloodsugar** and move it into the Dependent box.
3. Select the variable **Weight** and move it into the Independent(s) box, as illustrated in Figure 6.7.

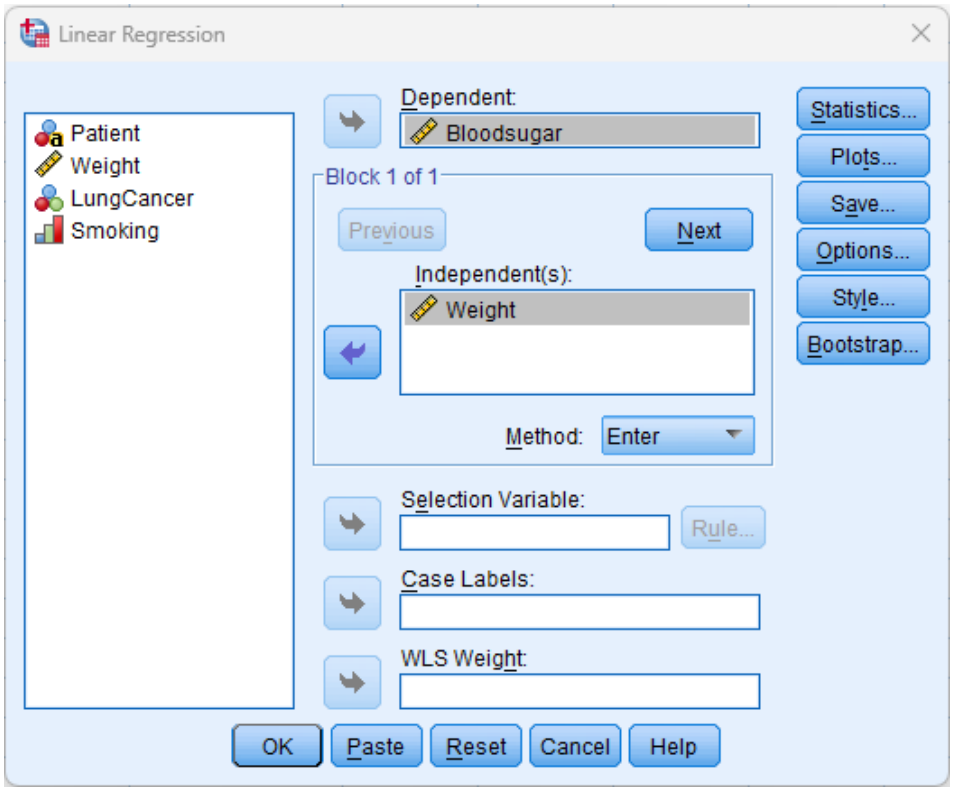


Figure 6.7: Linear Regression Dialog Box.

4. Click on **OK**: SPSS performs the linear regression (see Figure 6.8).

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Weight[b]	.	Enter

a. Dependent Variable: Bloodsugar

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,925[a]	,855	,837	,51969

a. Predictors: (Constant), Weight

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	12,767	1	12,767	47,270	,000[b]
	Residual	2,161	8	,270		
	Total	14,927	9			

a. Dependent Variable: Bloodsugar

b. Predictors: (Constant), Weight

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-,608	,405		-1,502	,172
	Weight	,027	,004	,925	6,875	,000

a. Dependent Variable: Bloodsugar

Figure 6.8: Linear Regression Output.

The B column contains the regression coefficients (a: the slope, b: the intercept) that you would use in a prediction equation. In this example, blood sugar level can be predicted using the following equation:

$$Bloodsugar = a * Weight + b$$

$$Bloodsugar = 0.027 * Weight - 0.608$$

6.5 Graphical Representation of Data

The **Graphs** menu in SPSS contains three main options: **Chart Builder**, **Graphboard Template Chooser**, and **Legacy Dialogs**. These options are different

ways of performing the same task. The **Legacy Dialogs** are the original SPSS charts and are mainly chosen by people who have used them for years and find it too difficult to switch to another option.

Graphboard Template Chooser and **Chart Builder** allow you to create charts in different ways. In the "**Graphboard Template Chooser**", you first select the variables you want to display. Based on this information, different chart options are suggested. **Chart Builder** begins by presenting different types of charts. After selecting a chart, you then specify the variables you will use.

6.5.1 Building Charts Using Chart Builder

SPSS includes **Chart Builder**, which uses a graphical display to guide you through the steps of constructing charts. The program continuously checks your actions and prevents the use of features that are not compatible or would not function properly.

Through an example, we will see how to generate a graph. Suppose we want to plot the **boxplot** of the variable "**Bloodsugar**".

1. Choose **Graphs ➜ Chart Builder**. A warning appears informing you that before using this dialog box, the measurement level must be correctly defined for each variable in your chart. (We have defined the correct measurement level, so you can continue.)
2. Click **OK**: the **Chart Builder** dialog box appears.
3. Make sure the **Gallery** tab is selected: in the "**Choose from**" list, select "**Boxplots**" as the chart type.
4. Different types of boxplots appear in the gallery to the right of the list, as illustrated in Figure 6.9 below.

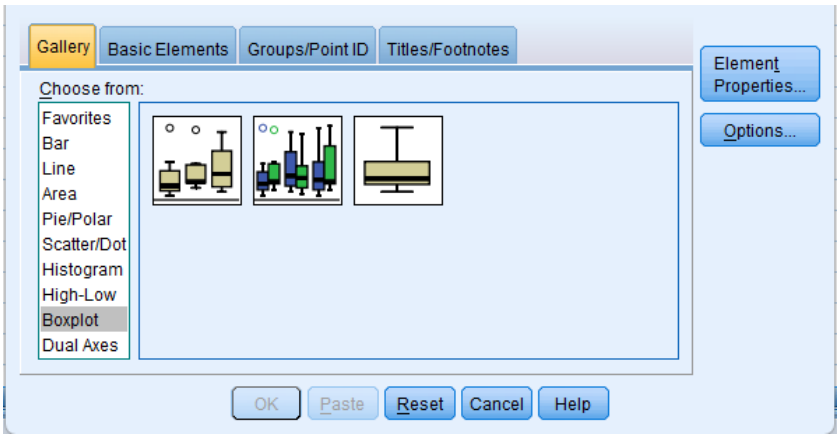


Figure 6.9: Chart Builder.

5. Select and drag "**1D: Boxplot**" into the canvas panel. The **Element Properties** tab now appears to the right of the preview panel at the top. This tab lets you know which features of the element you can modify. For example, you can change the statistic displayed or the chart style. In this

example, you will not use the **Element Properties** tab, so simply close it.

6. In the variable list, select the variable **Bloodsugar** and drag it to the Y-axis label in the chart. The graphical display inside the chart preview window never represents your actual data, even after inserting variable names.
7. Click the **OK** button to produce the chart: the resulting output is illustrated in Figure 6.10 below.

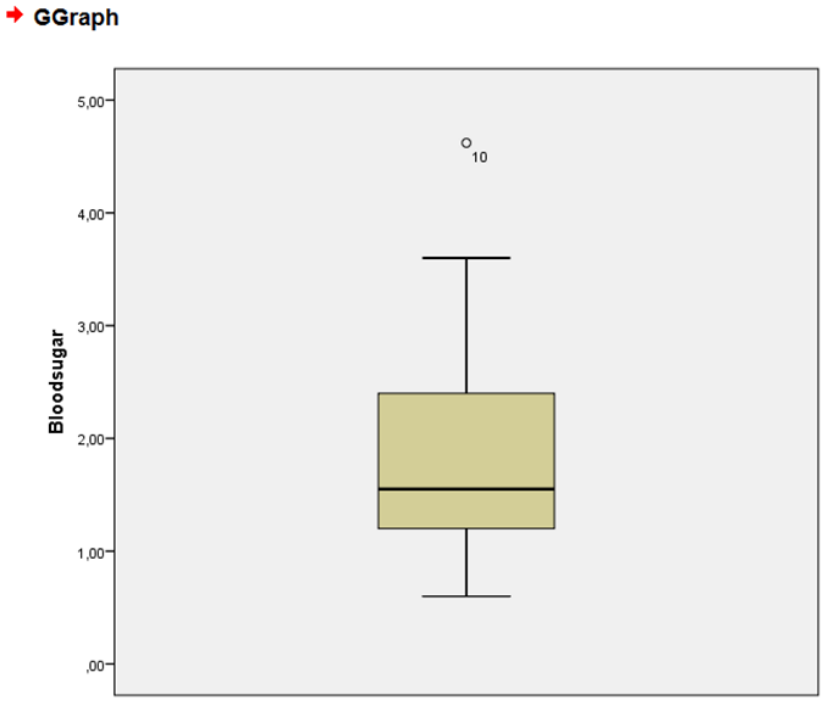


Figure 6.10: Boxplot: 1D for the variable Bloodsugar.

6.5.2 Displaying a Linear Relationship

The following steps show you how to construct a simple scatterplot:

1. Choose **Graphs ➜ Chart Builder**.
2. Click the **OK** button and then the **Reset** button.
3. In the "**Choose from**" list, select **Scatter/Dot**.
4. Select the first scatterplot (Simple Scatter) and drag it to the panel at the top.
5. In the Variables list, select **Weight** and drag it to the box labeled X-Axis in the chart.
6. In the Variables list, select **Bloodsugar** and drag it to the box labeled Y-Axis in the chart.
7. Click **OK**: the chart in Figure 6.11 is displayed.

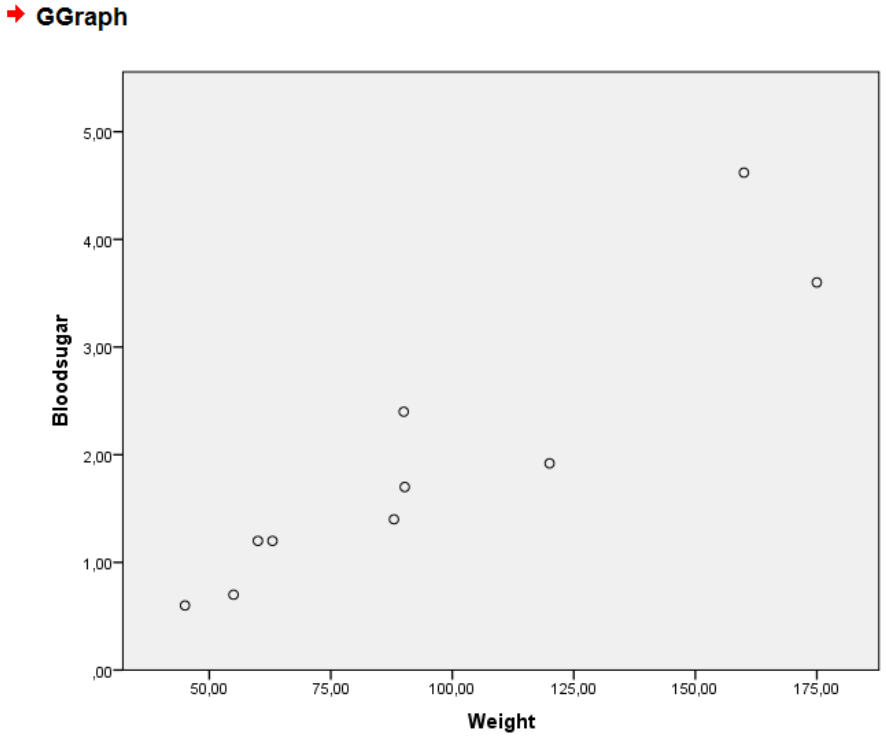


Figure 6.11: Scatterplot of Two Quantitative Variables

8. Double-click the produced chart: the "**Chart Editor**" dialog box appears.
9. Click the icon "**Add Fit Line at Total**", then click the Close button, and finally close the "**Chart Editor**" window.

10. The resulting output is illustrated in Figure 6.12 below.

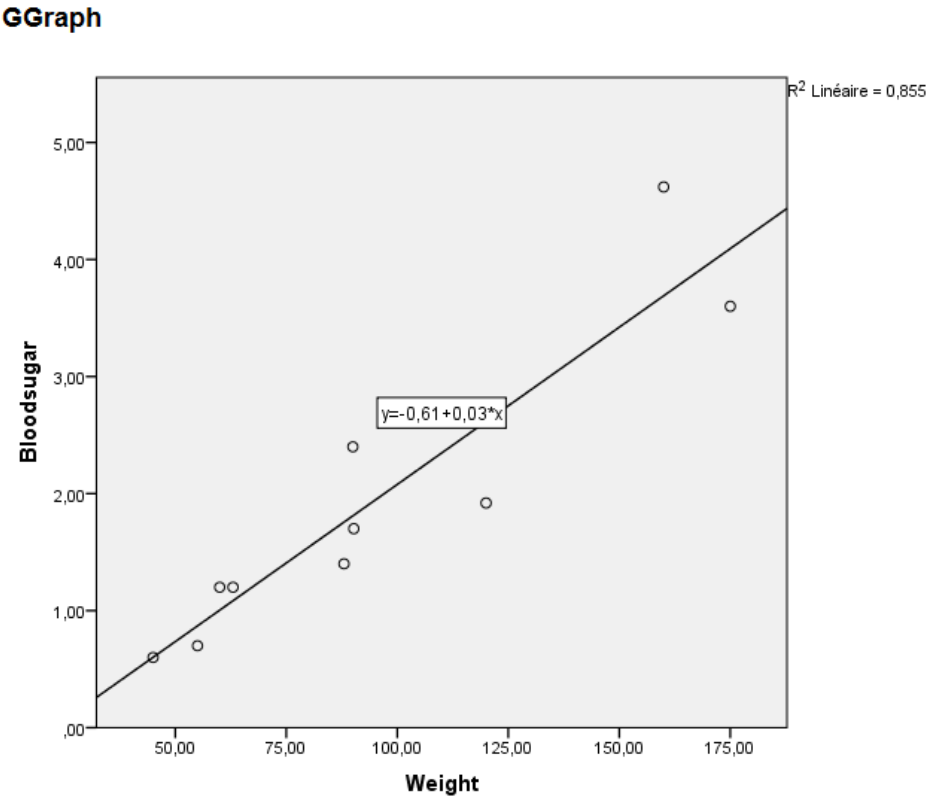


Figure 6.12: Linear Regression Line

The line superimposed on the scatterplot is the best-fitting line that describes the linear relationship given by the equation: **y = 0.03*x - 0.61**, where y represents **Bloodsugar** and x represents **Weight**.

6.6 Conclusion

Researchers can use SPSS to study relationships between variables by analyzing both qualitative and continuous variables. SPSS provides various tools and functions, such as cross-tabulation and correlation analysis, to examine the association or correlation between variables. Cross-tabulation analysis can be used to identify relationships between two or more qualitative variables, while correlation analysis can be used to determine the strength and direction of the relationship between two continuous variables. In addition, regression analysis can be used to examine the relationship between a dependent variable and one or more independent variables, and to create a model that predicts the values of the dependent variable based on the independent variables. By using these functions and tools, researchers can study the relationships between variables in their biomedical data and make informed decisions based on their analysis.

Practical Works

TP 1

Objective

In this practical session, we will learn how to create an Excel file containing multiple quantitative and qualitative variables. We will also focus on learning how to validate the entered data to ensure its accuracy.

Exercise

	A	B	C	D	E
1	Patient	Age	Sexe	Blood	Temperature
2	P1	77	Femme	B	37,4
3	P2	64	Homme	A	38
4	P3	87	Homme	AB	38,5
5	P4	70	Femme	A	38,2
6	P5	85	Femme	O	38,1

1. Create an Excel data file with different types of variables (qualitative and quantitative):
   (a) Create a new Excel data file.
   (b) The dataset should contain five variables named: Patient, Age, Sex, Blood, and Temperature (see the figure above).
   (c) Enter the data for the "Patient" variable from P1 to P5 using the fill handle.
2. Before entering data for the variables Age, Sex, Blood, and Temperature, ensure that each variable is properly validated:
   (a) Check the validity of quantitative variables (Age, Temperature).
   (b) Check the validity of qualitative variables (Sex, Blood).
3. Save the resulting Excel file as `Tp1.xlsx`.
4. Delete the variable "Temperature" and save the dataset in another file called `Tp1_2.xlsx`.

Solution

1. (a) Open Microsoft Excel.

   (b) In the first row, enter the names of the five variables in each column: A1: Patient, B1: Age, C1: Sex, D1: Blood, and E1: Temperature.

   (c) Select cell A2 ➜ Enter P1 ➜ drag the fill handle down to A6 to fill the patient IDs.

2. (a) For the discrete quantitative variable Age: select the cells B2:B6 ➜ Data ➜ Data Validation ➜ Choose Allow: Whole number ➜ Data: between ➜ Minimum: 1 ➜ Maximum: 120 ➜ OK.

   (b) For the continuous quantitative variable Temperature: select the cells E2:E6 ➜ Data ➜ Data Validation ➜ Choose Allow: Decimal ➜ Data: between ➜ Minimum: 29.5 ➜ Maximum: 42.3 ➜ OK.

   (c) For the qualitative variable Sex: select the cells C2:C6 ➜ Data ➜ Data Validation ➜ Choose Allow: List ➜ Source: Homme;Femme ➜ OK.

   (d) For the qualitative variable Blood: select the cells D2:D6 ➜ Data ➜ Data Validation ➜ Choose Allow: List ➜ Source: AB;A;B;O ➜ OK.

   (e) Fill in the data as shown in the figure above.

3. File Tab ➜ Save ➜ File name: `Tp1.xlsx` ➜ Save.

4. Select column E ➜ press the "Delete" key ➜ File Tab ➜ Save As ➜ File name: `Tp1_2.xlsx` ➜ Save.

For more details, refer to Chapter 2.

TP 2

Objective

In this practical session, we will learn how to enter data and correctly define variables using the SPSS program. In addition, it is important to know how to manage SPSS files and how to transfer data from Excel to SPSS.

Exercise

	Patient	Sexe	Temperature	Bloodsugar	Hypertension
1	P1	1	38,10	1,20	1
2	P2	1	37,10	1,70	2
3	P3	2	38,00	1,40	1
4	P4	1	38,10	,60	1

1. Create an SPSS data file with different types of variables (qualitative and quantitative):
   (a) Create a new SPSS data file called “Tp2.sav”.
   (b) The data must contain five variables named: Patient, Sex, Temperature, Bloodsugar, and Hypertension (see the figure above).
   (c) The variable “Patient” is a String-type variable.
   (d) Enter the possible values for the qualitative variable “Sex”: 1 = Male and 2 = Female.
   (e) The variable “Temperature” is a continuous quantitative variable of Numeric type.
   (f) The variable “Bloodsugar” is a continuous quantitative variable of Numeric type.
   (g) Enter the possible values for the qualitative variable “Hypertension”: 1 = Yes, 2 = No.
2. Save the resulting SPSS file.
3. Create another Excel file and try opening it from SPSS.

Solution

1. To enter data in SPSS, you can follow these steps:
   (a) Open the SPSS software and click on the “Variable View” tab.
   (b) Enter the variable names in the first column of the table. Each variable name must be unique and descriptive.

(c) Define the variable properties (metadata) in the columns: (Name, Type, Width, Decimals, Label, Values, Missing, Columns, Align, Measure).

(d) Go to the Data View and enter the patients' data.

2. Once you have entered all the data, save the file by clicking "Save" in the "File" menu.

3. You can also import data into SPSS from other software or file formats, such as Excel. To do this, click on:

   File ➜ Open ➜ Data ➜ Excel File Type (*.xls *.xlsx, *.xlsm) ➜ select the Excel file ➜ Open.

   After importing the data, you may need to modify the variable types or recode the data to meet your analysis needs.

For more details, refer to Chapter 3.

TP 3

Objective

1. Help the student understand the appropriate techniques for accurately collecting and entering data using the SPSS software.
2. This practical work aims to highlight best practices for data processing in SPSS, including strategies for handling missing data, sorting data, recoding variables, splitting data, and selecting observations.

Exercise 1

	Patient	Weight	Bloodsugar	Sexe	Bloodgroup
1	P1	53,00	1,20	1	1
2	P2	44,00	-99,00	1	3
3	P3	88,00	1,40	2	2
4	P4	45,00	,60	1	4
5	P5	90,00	2,40	2	2
6	P6	175,00	3,60	2	1

1. Create a new SPSS data file called "Tp3.sav".
2. The data must contain five variables named: Patient, Weight, Bloodsugar, Sexe, and Bloodgroup (see the figure above).
3. The variable "Patient" is a string variable.
4. The variable "Weight" is a continuous quantitative variable of numeric type.
5. The variable "Bloodsugar" is a continuous quantitative variable of numeric type. (For the Bloodsugar variable, missing values are represented by the number -99).
6. Enter the possible values for the qualitative variable "Sexe": 1=Homme and 2=Femme.
7. Enter the possible values for the qualitative variable "Bloodgroup": 1=AB , 2=A, 3=B and 4=O.

Exercise 2

1. In the variable Bloodsugar, replace the missing values with the mean value of the series?
2. Sort the patient data according to their Bloodsugar values?
3. In a new variable named "BloodsugarCat": try to recode the values of the variable "Bloodsugar" as follows: Class 1: Normal $\leq$ 0.99 ; Class 2: Prediabetes]0.99 - 1.25] ; Class 3: Diabetes > 1.25.

4. Delete the new variable "BloodsugarCat"?
5. Compare the patient data separately according to their Bloodgroup values?
6. Cancel the data splitting?
7. Select patients with a weight $\geq$ 70 and a Bloodsugar < 1 or male patients (Homme) with a Bloodgroup equal to AB?

Solution

1. Choose Transform ➜ Replace Missing Values ➜ Move the variable Bloodsugar to the "New Variables" area ➜ OK ➜ Manually replace the value -99 in the Bloodsugar column with 1.84 ➜ Delete the column Bloodsugar_1.
2. Choose Data ➜ Sort Cases ➜ Sort by: Bloodsugar ➜ OK.
3. Choose Transform ➜ Recode into Different Variables ➜ Move the variable Bloodsugar to the working area on the right ➜ In Name: name the new variable BloodsugarCat ➜ Change ➜ Old and New Values ➜ For category 1: select the radio button Range, from LOWEST through value ➜ Enter the Max value (i.e., 0.99) ➜ Next to the Value radio button: enter the category number (i.e., 1) ➜ Click the Add button ➜ For category 2: select the radio button Range ➜ Enter the Min (i.e., 0.99) and Max (i.e., 1.25) values ➜ Next to the Value radio button: enter the category number (i.e., 2) ➜ Click the Add button ➜ For category 3: select the radio button Range, from value through HIGHEST ➜ Enter the Min value (i.e., 1.25) ➜ Next to the Value radio button: enter the category number (i.e., 3) ➜ Click the Add button ➜ Continue ➜ OK.
4. In the Data View tab, click on the variable name "BloodsugarCat" ➜ Press the Delete key on the keyboard.
5. Choose Data ➜ Split File ➜ Select the radio button "Compare groups" ➜ Choose Bloodgroup as the grouping variable ➜ OK ➜ Choose Analyze ➜ Descriptive Statistics ➜ Frequencies ➜ Choose Sexe and move it to the Variable(s) area ➜ OK.
6. Choose Data ➜ Split File ➜ Reset ➜ OK.
7. Choose Data ➜ Select Cases ➜ Select the radio button "If condition is satisfied" ➜ Click the If... button ➜ In the expression box, write "(Weight >= 70 & Bloodsugar<1) | (Sexe = 1 & Bloodgroup = 1)" ➜ Continue ➜ OK.

For more details, refer to Chapter 4.

TP 4

Objective

1. Help the student understand the appropriate techniques for accurately collecting and entering data using the SPSS software.
2. Review some of the methods used in SPSS data processing.
3. Present the techniques used to analyze qualitative and quantitative data using frequency and descriptive procedures specifically in SPSS.

Exercise 1

	Patient	Weight	Bloodsugar	Sexe	Bloodgroup
1	P1	63,00	1,20	1	1
2	P2	90,20	1,70	1	3
3	P3	88,00	1,40	2	2
4	P4	45,00	,60	1	4
5	P5	90,00	2,40	2	2
6	P6	175,00	3,60	2	1
7	P7	63,00	2,10	1	2

1. Create a new SPSS data file called "Tp4.sav".
2. The data must contain five variables named: Patient, Weight, Bloodsugar, Sexe, and Bloodgroup (see the figure above).
3. The variable "Patient" is a string variable.
4. The variable "Weight" is a continuous quantitative variable of numeric type.
5. The variable "Bloodsugar" is a continuous quantitative variable of numeric type.
6. Enter the possible values for the qualitative variable "Sexe": 1=Homme and 2=Femme.
7. Enter the possible values for the qualitative variable "Bloodgroup": 1=AB , 2=A, 3=B and 4=O.

Exercise 2

1. Sort the patient data according to their Weight and Bloodsugar values. What do you notice? (Focus on patients P1 and P7).
2. Compare the patient data separately according to their Sexe values.
3. Based on the previous data splitting, we want to study the distribution of the variable Bloodgroup (AB, A, B, O) for each value of Sexe (Homme,

Femme). Statistically and graphically analyze the qualitative variable Bloodgroup using the frequency procedure.

4. Cancel the data splitting?
5. Statistically and graphically analyze the quantitative variable Weight using the frequency procedure.
6. Statistically summarize the continuous variables using the descriptive procedure.

Solution

1. Choose Data ➜ Sort Cases ➜ Sort by: Weight and Bloodsugar, respectively ➜ OK. (We observe that P1 is positioned before P7.)
2. Choose Data ➜ Split File ➜ Select the radio button "Compare groups" ➜ Choose Sexe as the grouping variable ➜ OK.
3. Choose Analyze ➜ Descriptive Statistics ➜ Frequencies ➜ place Bloodgroup in the Variable(s) area ➜ Statistics ➜ Check the Mode box ➜ Continue ➜ Charts ➜ Select the Bar charts radio button ➜ Select the Percentages radio button ➜ Continue ➜ OK.
4. Choose Data ➜ Split File ➜ Reset ➜ OK.
5. Choose Analyze ➜ Descriptive Statistics ➜ Frequencies ➜ Place the variable Weight in the Variable(s) area ➜ Statistics ➜ Check the following boxes: Mean, Median, Mode, Standard deviation, Variance, Minimum and Maximum ➜ Continue ➜ Charts ➜ Select the Histogram radio button ➜ Check the box "Show normal curve on histogram" ➜ Continue ➜ Uncheck the box "Display frequency tables" ➜ OK.
6. Choose Analyze ➜ Descriptive Statistics ➜ Descriptives ➜ Place the variables Weight and Bloodsugar in the Variable(s) area ➜ OK.

For more details, refer to Chapter 5.

TP 5

Objective

1. Help the student understand the appropriate techniques for accurately collecting and entering data using the SPSS software.
2. To show how to analyze relationships between qualitative variables using cross-tabulations: Cross-tabulations are used to create contingency tables that summarize the relationship between two or more categorical variables. This practical work aims to show how to create and interpret these tables in SPSS.
3. Demonstrate how to analyze relationships between quantitative variables using correlation and linear regression: Correlation and linear regression are statistical methods used to model the relationship between two or more continuous variables. This practical work aims to show how to use these methods to create and interpret regression models in SPSS.
4. Demonstrate how to create plots and charts in SPSS: This practical work aims to show how to use SPSS to create data visualizations in order to better understand relationships between variables or to communicate results to others.

Exercise 1

	Patient	Weight	Bloodsugar	Sexe	Bloodgroup
1	P1	63,00	1,20	1	1
2	P2	90,20	1,70	1	3
3	P3	88,00	1,40	2	2
4	P4	45,00	,60	1	4
5	P5	90,00	2,40	2	2
6	P6	175,00	3,60	2	1
7	P7	60,00	1,20	1	2
8	P8	120,00	1,92	1	1
9	P9	55,00	,70	2	4
10	P10	160,00	4,62	2	3

1. Create a new SPSS data file called "Tp5.sav".
2. The data must contain five variables named: Patient, Weight, Bloodsugar, Sexe, and Bloodgroup (see the figure above).
3. The variable "Patient" is a string variable.
4. The variable "Weight" is a continuous quantitative variable of numeric type.
5. The variable "Bloodsugar" is a continuous quantitative variable of numeric type.

6. Enter the possible values for the qualitative variable "Sexe": 1=Homme and 2=Femme.
7. Enter the possible values for the qualitative variable "Bloodgroup": 1=AB , 2=A, 3=B and 4=O.

Exercise 2

1. Sort the patient data according to their Weight values.
2. Using cross-tabulations, is there a relationship between the qualitative variables Sexe and Bloodgroup?
3. Prove that there is a linear correlation between the variables Bloodsugar and Weight. If yes, try to extract this linear equation using the linear regression technique.
4. Graphically represent the linear relationship between the variables Bloodsugar and Weight.

Solution

1. Choose Data ➜ Sort Cases ➜ Sort by: Weight ➜ OK.
2. Choose Analyze ➜ Descriptive Statistics ➜ Crosstabs ➜ Place the variable Sexe in the Row(s) area and the variable Bloodgroup in the Column(s) area ➜ Cells ➜ Select Row percentages ➜ Continue ➜ OK.
3. (a) Choose Analyze ➜ Correlate ➜ Bivariate ➜ Place the variables Weight and Bloodsugar in the Variables area ➜ Options ➜ Check the option "Cross-product deviations and covariances" ➜ Continue ➜ OK.

   (b) Select Analyze ➜ Regression ➜ Linear ➜ Place the variable Bloodsugar in the Dependent area and the variable Weight in the Independent(s) area ➜ OK.
4. Choose Graphs ➜ Chart Builder ➜ OK ➜ Reset ➜ Scatter/Dot ➜ Simple Scatter ➜ Select Weight and drag it to the rectangle labeled X-Axis in the chart ➜ Select Bloodsugar and drag it to the rectangle labeled Y-Axis in the chart ➜ OK ➜ Double-click the generated chart ➜ Click the icon "Add Fit Line at Total" ➜ Close ➜ Close the "Chart Editor" window.

For more details, refer to Chapter 6.

Bibliographical References

[Cronk, 2019] CRONK, B. C. (2019). *How to use SPSS®: A step-by-step guide to analysis and interpretation*. Routledge.

[Denis, 2018] DENIS, D. J. (2018). *SPSS data analysis for univariate, bivariate, and multivariate statistics*. John Wiley & Sons.

[Jean-Pierre, 2016] JEAN-PIERRE, L. (2016). Statistiques et probabilités: Cours et exercices corrigés.

[Microsoft Corporation,] MICROSOFT CORPORATION. Microsoft excel. https://support.microsoft.com/en-us/excel. Accessed: 2019-09-01.

[Nasir *et al.*, 2022] NASIR, M. A., BAKOUCH, H. S. et JAMAL, F. (2022). *Introductory Statistical Procedures with SPSS*. Bentham Science Publishers.

[Salcedo et McCormick, 2020] SALCEDO, J. et MCCORMICK, K. (2020). *SPSS Statistics for Dummies*. John Wiley & Sons.

[Twigg, 2010] TWIGG, M. (2010). Discovering statistics using spss.

[uiowa, 2024] UIOWA (2024). Normal distribution applet/calculator. https://homepage.divms.uiowa.edu/~mbognar/applets/normal.html. Accessed: 2024-02-01.