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 : November 30, 2025

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):

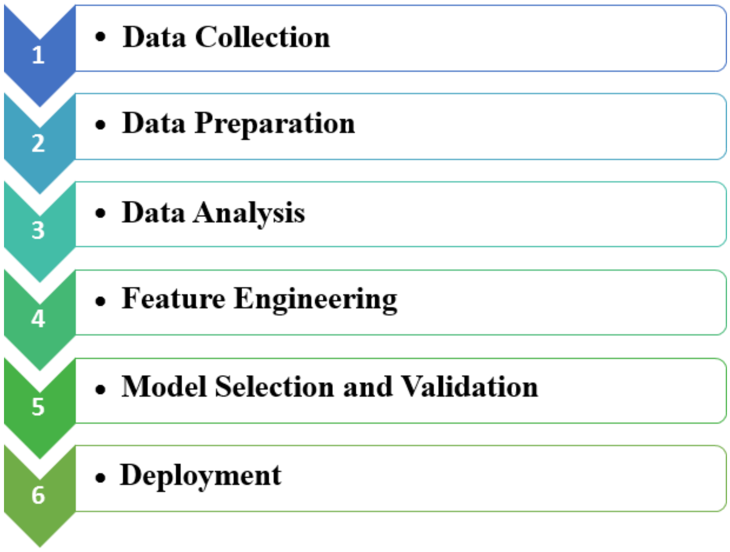


Figure 1.1: Flowchart of the Data Science Methodology

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).

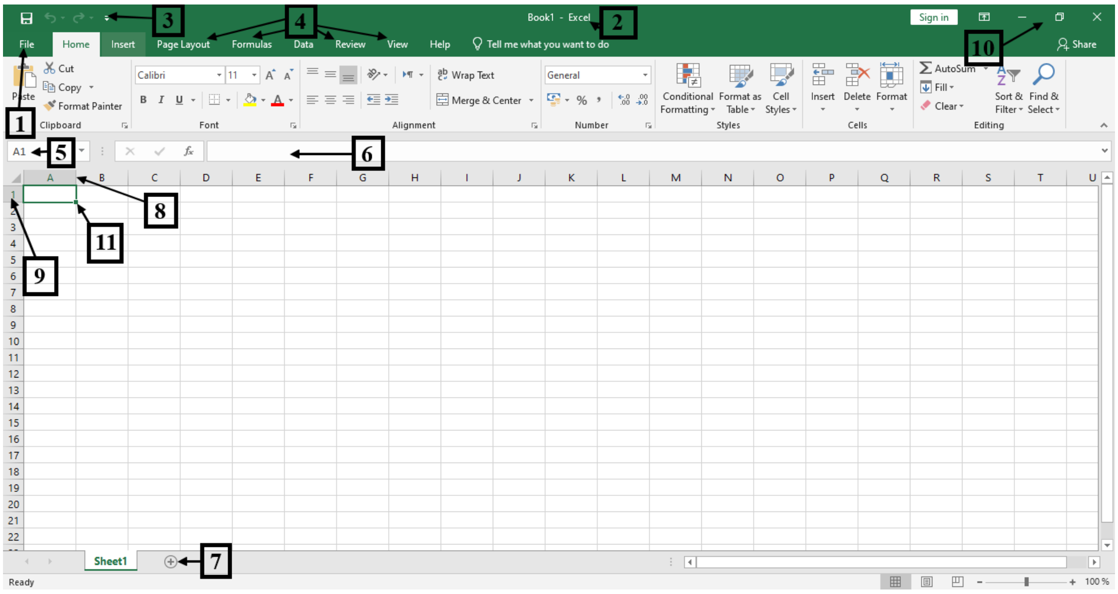


Figure 2.1: Excel Environment.

- 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.

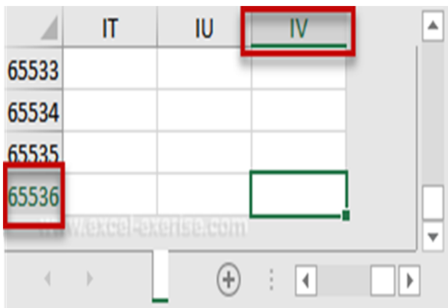
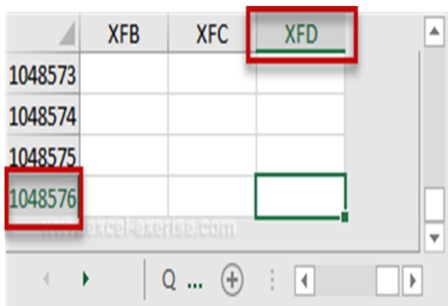
Excel 97-2003	**Excel 2007 and later (2010, 2013, 2016, 2019, 2021, 365)**
In an XLS workbook, the limits are 65,536 rows and 256 columns, which corresponds to column IV.	In an XLSX workbook, the limits are 1,048,576 rows and 16,384 columns, which corresponds to column XFD.
	

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).

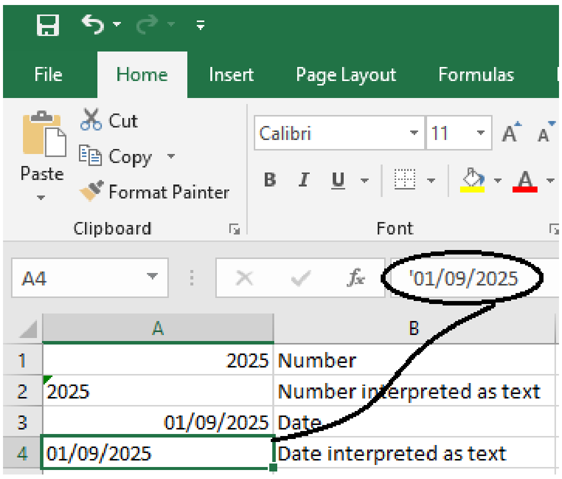


Figure 2.2: Using the apostrophe to enter text in Excel.

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.

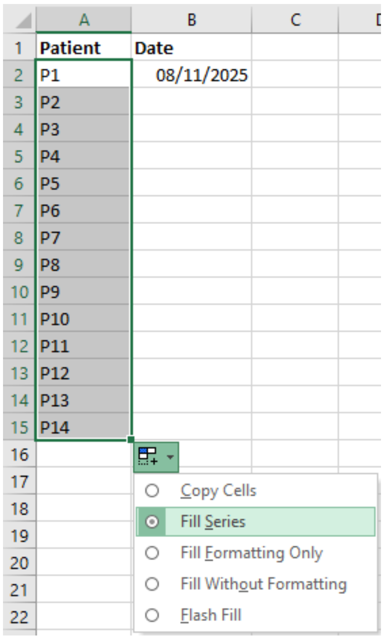


Figure 2.3: Using the fill handle in Excel to complete a data series.

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.

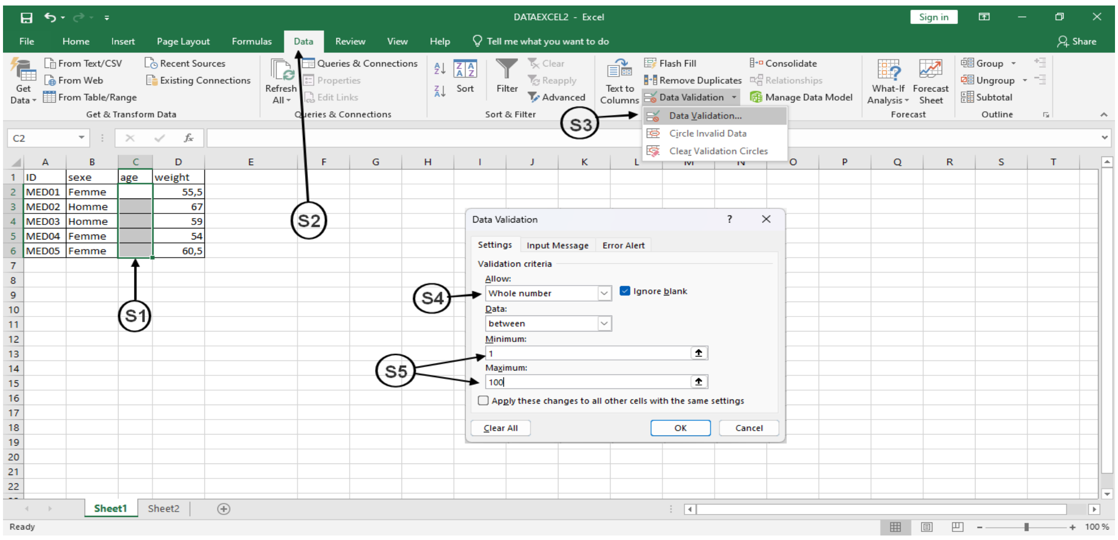


Figure 2.4: Data validation for quantitative variables

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).

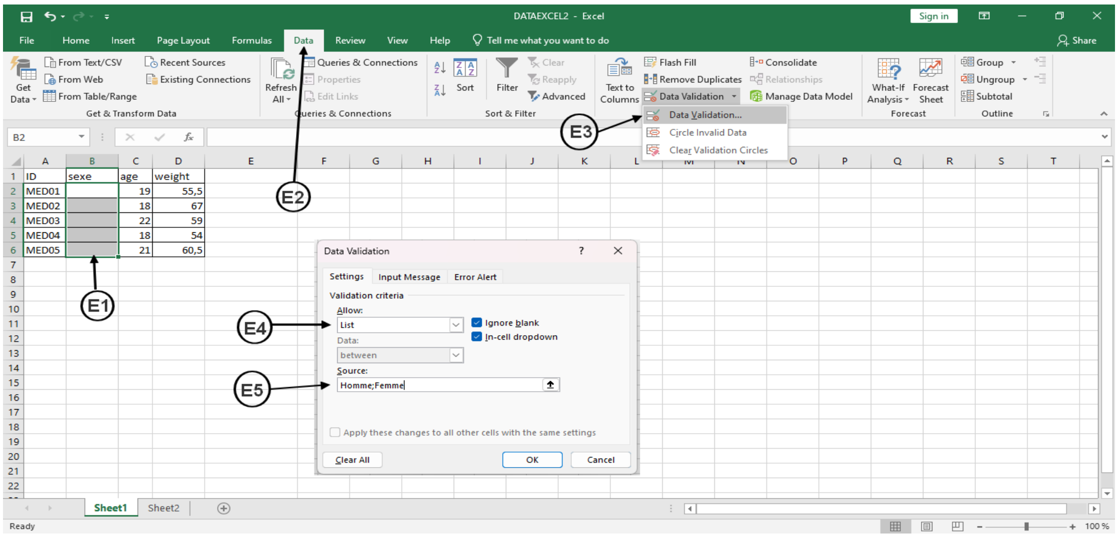


Figure 2.5: Data validation for qualitative variables

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:

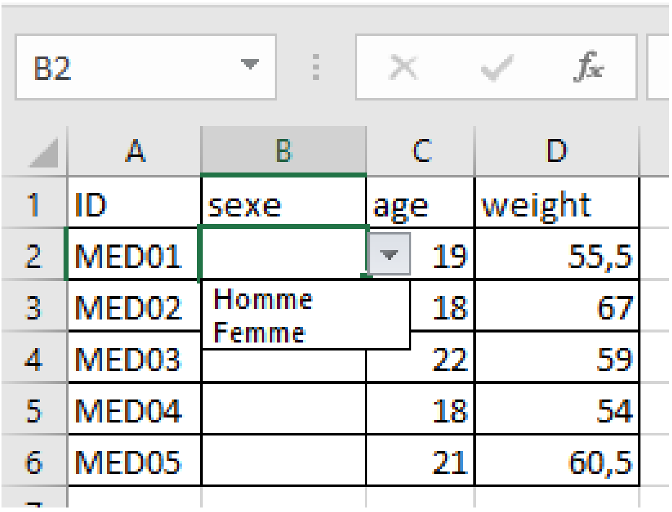


Figure 2.6: Result of data validation for qualitative variables

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.

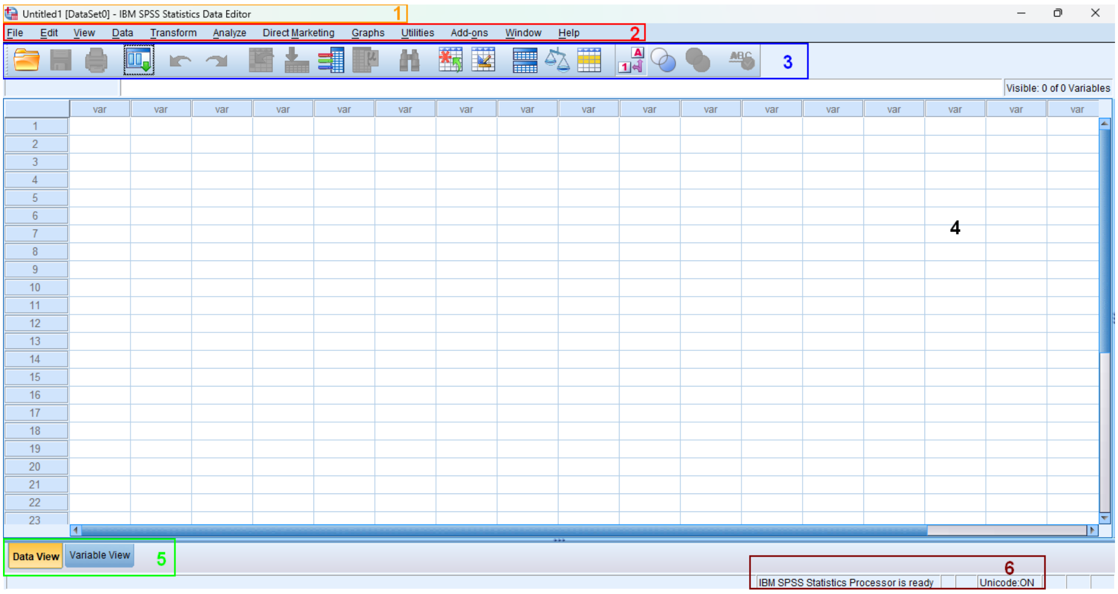


Figure 3.1: SPSS Environment.

- **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.

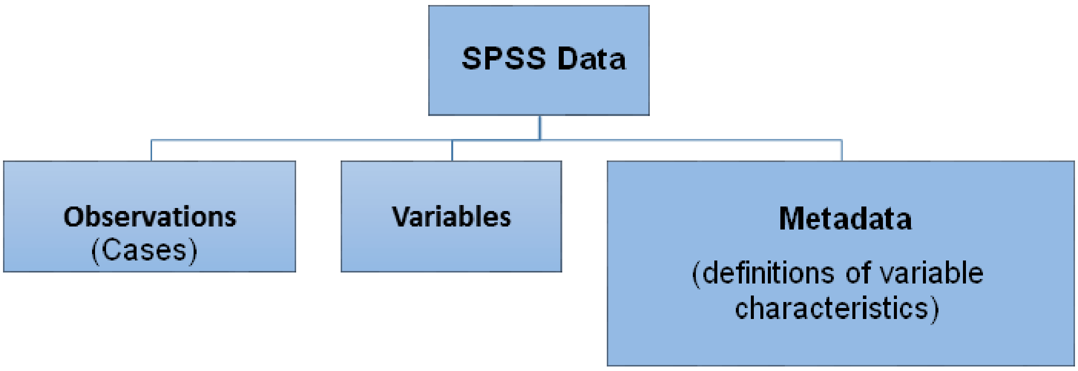


Figure 3.2: SPSS data.

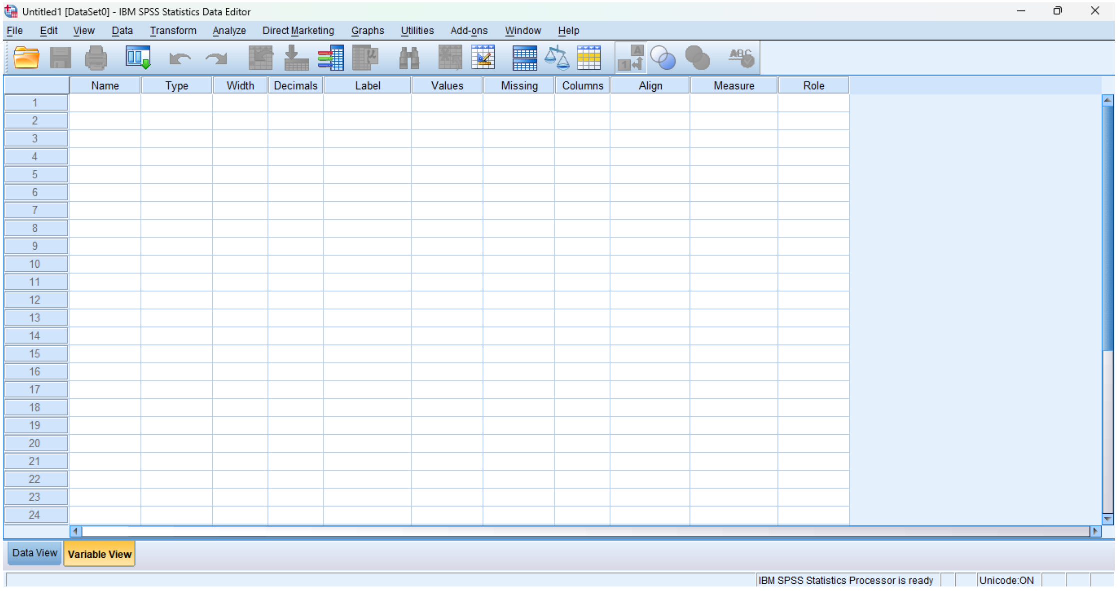


Figure 3.3: Variable View.

■ 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 (_).

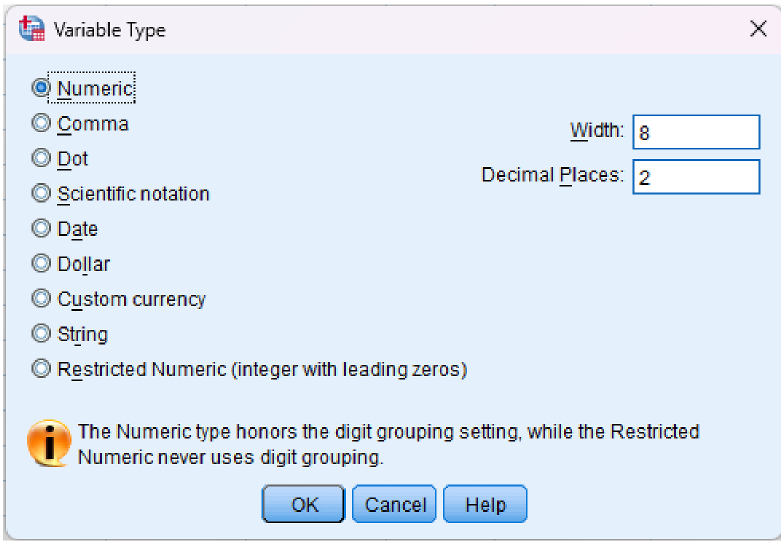


Figure 3.4: Variable type.

- 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.

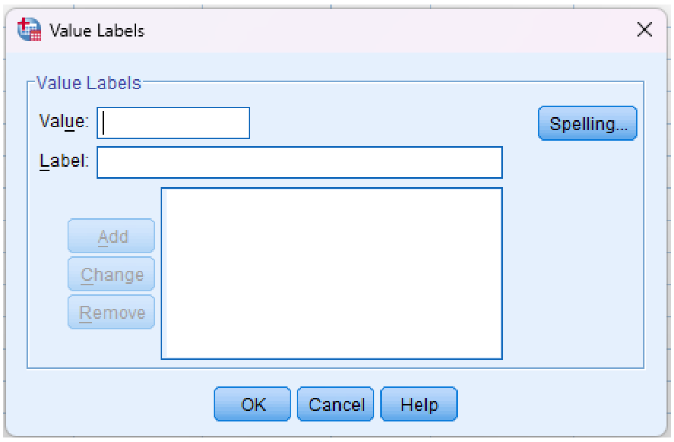


Figure 3.5: Value Labels.

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.

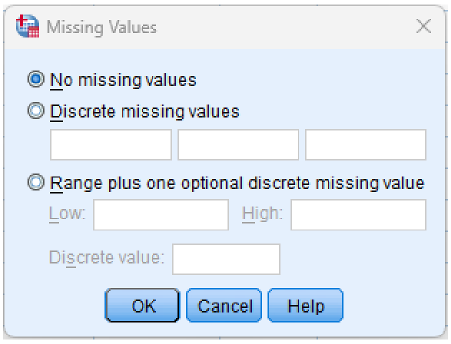


Figure 3.6: Missing values.

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).

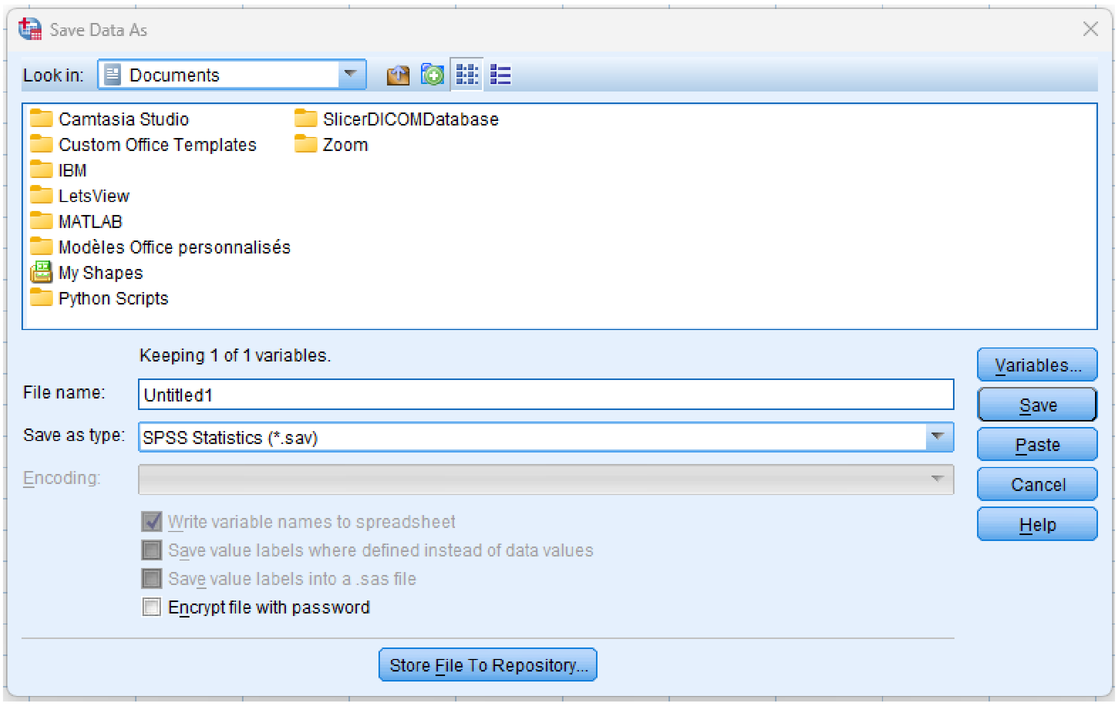


Figure 3.7: Saving Data in SPSS.

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.

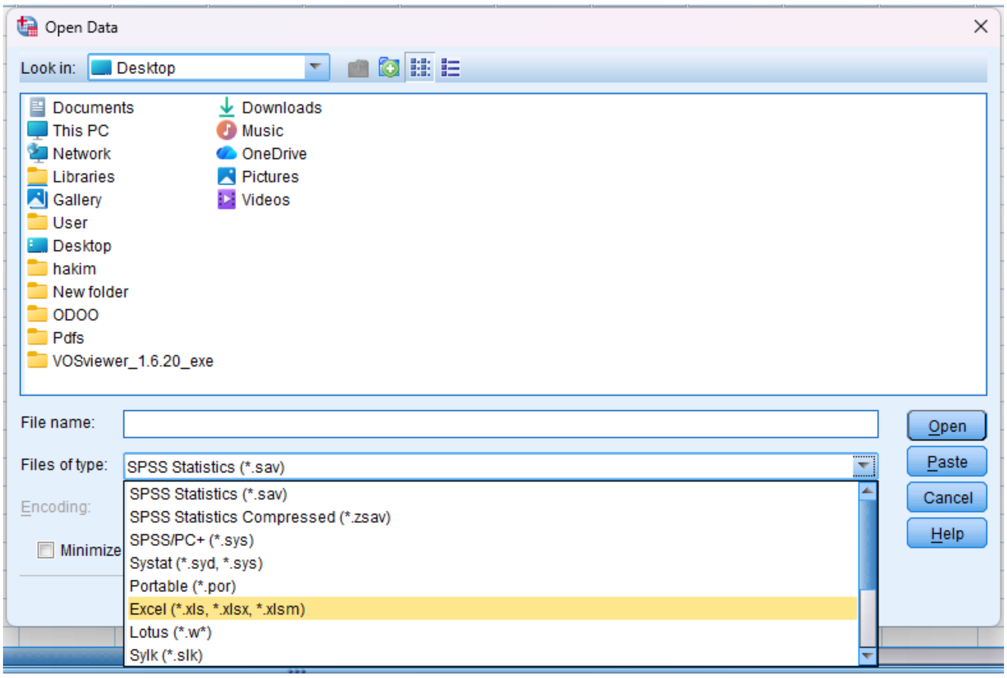


Figure 3.8: Opening Data in SPSS.

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).

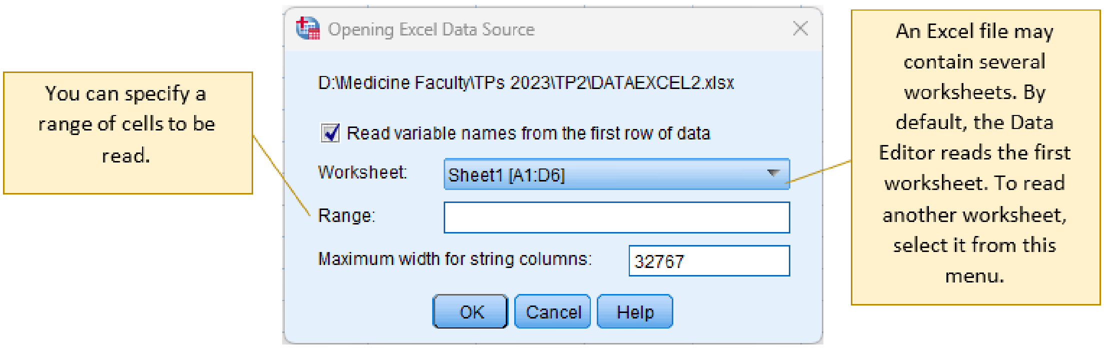


Figure 3.9: Open the Excel Data Source.

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.

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.

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.