Introduction

A large number of the actions performed by means of statistical software are essentially forms of manipulating, or even literally transforming digital data representing statistical data. It is therefore paramount to fully understand how statistical data are represented and how they can be employed by software such as Stata. After the importing, recoding and the eventual transformation of these data, the description of the variables of interest and the summary of their distribution in numerical and graphical form constitute a fundamental preparatory stage to any statistical modeling, hence the importance of these early stages in the progress of a project for statistical analysis. In a second step, it is essential to fully control the commands that enable the calculation of the main measures of association in medical research, and to know how to implement the conventional explanatory and predictive models: analysis of variance, linear and logistic regression and the Cox model. With a few exceptions, making use of the Stata commands available during the installation of the software (base commands) will be preferred over the usage of specialized libraries of commands.

This book assumes that the reader is already familiar with basic statistical concepts, in particular the calculation of central tendency and dispersion indicators for a continuous variable, contingency tables, analysis of variance and conventional regression models. The objective here is to apply this knowledge to datasets described in numerous other works, even if the interpretation of the results remains minimal, in order to quickly familiarize oneself with the use of Stata with actual data. Emphasis is particularly given to the management and the manipulation of structured data since it can be noted that this constitutes 60–80% of the work of the statistician. There are many books in French or in English on Stata, covering both the technical and the statistical point of view. Some of these works show a dominant generalistic nature [ACO 14, HAM 13, RAB 04], while others are much more specialized and address similar topics, such as [FRY 14, DUP 09, VIT 05]. The purpose of this book is to enable the reader to quickly become accustomed to Stata, so that they can

perform their own analyses and continue learning in an autonomous way in the field of medical statistics.

This book constitutes a sequel to the book *Biostatistics and Computer Analysis of Health Data using R* [LAL 16], published by the same authors in the same collection. Every topic that relates to data organization and data exploratory analysis, in particular graphical methods, are discussed therein. In this book, the same data sets are being used to facilitate the transfer of learning of the knowledge acquired in R.

In Chapter 1, the base commands for data management with Stata will be introduced. This primarily concerns the creation and the manipulation of quantitative and qualitative variables (recoding of individual values, counting of missing observations), importing databases stored in the form of text files, as well as elementary arithmetic operations (minimum, maximum, arithmetic mean, difference, frequency, etc.). We will also examine how to store preprocessed databases in text or in Stata formats. The objective is to understand how data are represented in Stata and how to work with them. The useful commands for describing a data table composed of quantitative or qualitative variables are also presented. The descriptive approach is strictly univariate, which constitutes the prerequisite for any statistical approach. Base graphic commands (histograms, density curves, bar or dot plots) will be presented in addition to the usual central tendency (mean, median) and dispersion (variance, quartiles) numerical descriptive summaries. Pointwise and interval estimation using arithmetic means and empirical proportions will also be addressed. The objective is to become familiar with the use of simple Stata commands operating on a variable, optionally specifying certain options for the calculation, alongside the selection of statistical units among all of the available observations.

Chapter 2 is dedicated to the comparison of two samples for quantitative or qualitative measurements. The following hypothesis tests are addressed: the Student's test for independent or paired samples, the non-parametric Wilcoxon test, the χ^2 test and the Fisher's exact test, and the NcNemar test based on the main measures of association for two variables (average difference, odds ratio and relative risk). From this chapter onwards, there will be less emphasis on the univariate description of each variable, but it is advisable to always carry out the stages of data description discussed in this chapter. The objective is to control the main statistical tests in the case where the relationship between a quantitative variable and a qualitative variable, or for two qualitative variables, is the main interest. This chapter also presents analysis of variance (ANOVA) where we explain the variability observed at the level of a numerical response variable by taking a group or classification factor into account, and the estimation with confidence intervals of average differences. Emphasis will be placed on the construction of an ANOVA table summarizing the various sources of variability, and on the graphic methods that can be used to summarize the distribution of individual or aggregated data. The linear tendency test will also be studied when the classification factor can be considered as

naturally ordered. The objective is to understand how to construct an explanatory model in the case where there is one or even two explanatory factors, and how to digitally and graphically present the results of such a model through the use of Stata.

Chapter 3 focuses on the analysis of the linear relation between two continuous quantitative variables. In the linear correlation approach, which assumes a symmetrical relation between the two variables, the main focus will be on quantifying the force and the direction of the association in a parametric (Pearson correlation) or in a non-parametric manner (rank-based Spearman correlation) and on the graphic representation of this relation. Simple linear regression will be used in the event that one of the two numeric variables assumes the function of a response variable, and the other that of an explanatory variable. The useful commands for the estimation of the coefficients of the regression line, the construction of the ANOVA table associated with the regression and the computation of fitted values will be presented. The objective of this chapter remains identical to that of Chapter 2, namely to present the Stata commands necessary for the construction of a simple statistical model between two variables following an explanatory or predictive perspective.

In Chapter 4, the main measures of association found in epidemiological studies will be discussed: odds ratio, relative risk, prevalence, etc. Stata commands allowing the estimation (pointwise and by interval) and the associated hypothesis tests will be illustrated with data from cohort or case–control studies. The implementation of a simple logistic regression model makes it possible to complete the range of statistical methods, allowing the observed variability to be explained at the level of binary response variables. The objective is to understand the Stata commands to be used in the case in which the variables are binary, either to summarize a contingency table in the form of association indicators or to model the relationship between a binary response (ill/healthy) and a qualitative explanatory variable based on the so-called grouped data.

Chapter 5 constitutes an introduction to the analysis of censored data, the main tests associated with the construction of a survival curve (log-rank or Wilcoxon tests) and finally the Cox regression model. The specificity of the censored data requires particular care in the coding of data in Stata, and the objective is to present the Stata commands essential to the correct representation of survival data in digital form, to their numerical (survival median) and graphical (Kaplan–Meier curve) summary, and the implementation of common tests.

At the end of each chapter, a few applications are provided and a few examples of commands that can be used to respond to most of the presented questions are proposed. It is sometimes possible to obtain identical results with other approaches or by utilizing other commands. Stata outputs are not reproduced but readers are encouraged to try themselves the proposed Stata instructions and to try alternative or complementary instructions. It will be assumed that the data files used are available

in the working directory. All of the data files and the Stata commands used in this book can be downloaded from the companion website (https://github.com/biostatsante).

Due to layout reasons, some of the Stata outputs have been truncated or reformatted. As a result, these could present differences when the reader attempts to reproduce the commands mentioned in this book.

An index of the Stata commands used in the illustrations is available at the end of the book.